DOCUMENT RESUME

ED 440 153 TM 030 841

AUTHOR	Horkay, Nancy, Ed.
TITLE	The NAEP Guide: A Description of the Content and Methods of the 1999 and 2000 Assessments. Revised Edition.
INSTITUTION	National Center for Education Statistics (ED), Washington, DC.
REPORT NO	NCES-2000-456
PUB DATE	1999-11-00
NOTE	86p.; For the 1997-1998 guide, see ED 412 294.
AVAILABLE FROM	ED Pubs, P.O. Box 1398, Jessup, MD 20794-1398. Tel: 877-433-7827 (Toll Free); Fax: 301-470-1244; Web site: http://www.ed.gov/pubs/edpubs.html. For full text: http://nces.ed.gov/nationsreportcard.
PUB TYPE	Reports - Descriptive (141)
EDRS PRICE	MF01/PC04 Plus Postage.
DESCRIPTORS	*Academic Achievement; Comparative Analysis; *Data Collection; Educational Assessment; Elementary Secondary Education; *National Competency Tests; National Surveys; *Research Methodology; Sampling; Scoring; State Surveys; *Test Bias; Test Interpretation; Testing Problems
IDENTIFIERS	*National Assessment of Educational Progress

ABSTRACT

As mandated by Congress, the National Assessment of Educational Progress (NAEP) surveys the educational accomplishments of U.S. students and monitors changes in those accomplishments. NAEP tracks the educational achievement of 4th, 8th, and 12th graders over time in selected content areas, and has been tracking achievement for 30 years. The goals of this guide are to provide readers with an overview of the project and to help them understand the philosophical approach, procedures, analysis, and psychometric underpinnings of the NAEP. It introduces NAEP resources, demonstrates how the survey design matches the NAEP role as an indicator of national achievement, and describes some of the methods used in the 1999 and 2000 assessments. The guide follows a question-and-answer format, presenting questions most commonly asked about the NAEP and following them with answers and background information. Questions relate to: (1) the nature of the NAEP; (2) the subjects it assesses; (3) participation and confidentiality; (4) the availability of information to parents and the public; (5) the evaluation and validation of the NAEP; (6) the assessment development process; (7) accommodations for students with disabilities and those of limited English proficiency; (8) assessment innovations; (9) results; (10) scoring; (11) analysis of results; (12) comparability of results; (13) NAEP reports; (14) contextual background data; (15) use of NAEP results by educators; (16) use of NAEP results in education and policy issues; (17) linkage with other assessments; (18) students assessed in the NAEP; (19) student and school participation; (20) use of matrix sampling; and (21) data collection procedures. (Contains 32 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

TM

NATIONAL CENTER FOR EDUCATION STATISTICS

ED 440 153

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☐ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

The NAEP Guide

TM030841

BEST COPY AVAILABLE

THE NATION'S REPORT CARD
U.S. Department of Education
Office of Educational Research and Improvement
NCES 2000-456

What Is The Nation's Report Card?

THE NATION'S REPORT CARD, the National Assessment of Educational Progress (NAEP), is the only nationally representative and continuing assessment of what America's students know and can do in various subject areas. Since 1969, assessments have been conducted periodically in reading, mathematics, science, writing, history, geography, and other fields. By making objective information on student performance available to policymakers at the national, state, and local levels, NAEP is an integral part of our nation's evaluation of the condition and progress of education. Only information related to academic achievement is collected under this program. NAEP guarantees the privacy of individual students and their families.

NAEP is a congressionally mandated project of the National Center for Education Statistics, the U.S. Department of Education. The Commissioner of Education Statistics is responsible, by law, for carrying out the NAEP project through competitive awards to qualified organizations. NAEP reports directly to the Commissioner, who is also responsible for providing continuing reviews, including validation studies and solicitation of public comment, on NAEP's conduct and usefulness.

In 1988, Congress established the National Assessment Governing Board (NAGB) to formulate policy guidelines for NAEP. The Board is responsible for selecting the subject areas to be assessed from among those included in the National Education Goals; for setting appropriate student performance levels; for developing assessment objectives and test specifications through a national consensus approach; for designing the assessment methodology; for developing guidelines for reporting and disseminating NAEP results; for developing standards and procedures for interstate, regional, and national comparisons; for determining the appropriateness of test items and ensuring they are free from bias; and for taking actions to improve the form and use of the National Assessment.

The National Assessment Governing Board

Mark D. Musick, Chair
President
Southern Regional Education Board
Atlanta, Georgia

Michael T. Nettles, Vice Chair
Professor of Education and Public Policy
University of Michigan
Ann Arbor, Michigan

Moses Barnes
Secondary School Principal
Fort Lauderdale, Florida

Melanie A. Campbell
Fourth-Grade Teacher
Topeka, Kansas

Honorable Wilmer S. Cody
Commissioner of Education
State of Kentucky
Frankfort, Kentucky

Edward Donley
Former Chairman
Air Products & Chemicals, Inc.
Allentown, Pennsylvania

Honorable John M. Engler
Governor of Michigan
Lansing, Michigan

Thomas H. Fisher
Director, Student Assessment Services
Florida Department of Education
Tallahassee, Florida

Michael J. Guerra
Executive Director
National Catholic Education Association
Secondary School Department
Washington, DC

Edward H. Haertel
Professor, School of Education
Stanford University
Stanford, California

Juanita Haugen
Local School Board President
Pleasanton, California

Honorable Nancy Kopp
Maryland House of Delegates
Bethesda, Maryland

Honorable William J. Moloney
Commissioner of Education
State of Colorado
Denver, Colorado

Mitsugi Nakashima
President
Hawaii State Board of Education
Honolulu, Hawaii

Debra Paulson
Eighth-Grade Mathematics Teacher
El Paso, Texas

Honorable Norma Paulus
Former Superintendent of Public Instruction
Oregon State Department of Education
Salem, Oregon

Honorable Jo Ann Pottorff
Kansas House of Representatives
Wichita, Kansas

Diane Ravitch
Senior Research Scholar
New York University
New York, New York

Honorable Roy Romer
Former Governor of Colorado
Denver, Colorado

John H. Stevens
Executive Director
Texas Business and Education Coalition
Austin, Texas

Adam Urbanski
President
Rochester Teachers Association
Rochester, New York

Deborah Voltz
Assistant Professor
Department of Special Education
University of Louisville
Louisville, Kentucky

Marilyn A. Whirry
Twelfth-Grade English Teacher
Manhattan Beach, California

Dennie Palmer Wolf
Senior Research Associate
Harvard Graduate School of Education
Cambridge, Massachusetts

C. Kent McGuire (Ex-Officio)
Assistant Secretary of Education
Office of Educational Research and Improvement
U.S. Department of Education
Washington, DC

Roy Truby
Executive Director, NAGB
Washington, DC

The NAEP Guide

A Description of the Content and Methods of the 1999 and 2000 Assessments

Revised Edition
November 1999

THE NATIONAL CENTER FOR EDUCATION STATISTICS
Office of Educational Research and Improvement
U.S. Department of Education

U.S. Department of Education
Richard W. Riley
Secretary

Office of Educational Research and Improvement
C. Kent McGuire
Assistant Secretary

National Center for Education Statistics
Gary W. Phillips
Acting Commissioner

Education Assessment Group
Peggy G. Carr
Associate Commissioner

November 1999

SUGGESTED CITATION:

U.S. Department of Education. National Center for Education Statistics. *The NAEP Guide*, NCES 2000–456, by Horkay, N., editor. Washington, DC: 1999.

FOR MORE INFORMATION:

To obtain single copies of this report, while supplies last, or ordering information on other U.S. Department of Education products, call toll free 1-877-4ED PUBS (877–433–7827), or write:

Education Publications Center (ED Pubs)
U.S. Department of Education
P.O. Box 1398
Jessup, MD, 20794–1398

TTY/TDD 1–877–576–7734
FAX 301–470–1244

Online ordering via the Internet: *http://www.ed.gov/pubs/edpubs.html*
Copies also are available in alternate formats upon request.
This report is also available on the World Wide Web: *http://nces.ed.gov/nationsreportcard*

Cover photo copyright 1999, PhotoDisc, Inc.

The work upon which this publication is based was performed for the National Center for Education Statistics, Office of Educational Research and Improvement, by Educational Testing Service.

ACKNOWLEDGMENTS

This guide was produced with the assistance of professional staff at the National Center for Education Statistics (NCES), Educational Testing Service (ETS), Aspen Systems Corporation, National Computer Systems (NCS), and Westat.

The NCES staff whose invaluable assistance provided text and reviews for this guide include: Janis Brown, Pat Dabbs, and Andrew Kolstad.

Many thanks are due to Nancy Horkay, who edited and coordinated production of the guide, and to the numerous reviewers at ETS. The comments and critical feedback of the following reviewers are reflected in this guide: Nancy Allen, Jay Campbell, Patricia Donahue, Jeff Haberstroh, Debra Kline, John Mazzeo, and Christine O'Sullivan. Thanks also to Connie Smith at NCS and Dianne Walsh at Westat for coordinating the external reviews.

The guide design and production were skillfully completed by Aspen staff members Wendy Caron, Robert Lee, John Libby, Laura Mitchell, Munira Mwalimu, Maggie Pallas, Amy Salsbury, and Donna Troisi.

NCES and ETS are grateful to Nancy Horkay, coordinator of the previous guide, upon which the current edition is based.

Introduction 1

Background and Purpose 3

Question 1:
What is NAEP? 3

Question 2:
What subjects does NAEP assess? How are the subjects chosen, and how are the assessment questions determined? What subjects were assessed in 1999? What subjects will be assessed in 2000? 6

Question 3:
Is participation in NAEP voluntary? Are the data confidential? Are students' names or other identifiers available? 14

Question 4:
Can parents examine the questions NAEP uses to assess student achievement? Can parents find out how well their children performed in the NAEP assessment? Why are NAEP questions kept confidential? 16

Question 5:
Who evaluates and validates NAEP? 18

Assessment Development 21

Question 6:
What process is used to develop the assessments? 21

Question 7:
How does NAEP accommodate students with disabilities and students with limited English proficiency? 24

Question 8:
What assessment innovations has NAEP developed? 26

Scoring and Reporting 28

Question 9:
What results does NAEP provide? 28

Question 10:
How does NAEP reliably score and process millions of student-composed responses? 31

Question 11:
How does NAEP analyze the assessment results? 35

Question 12:
How does NAEP ensure the comparability of results among the state assessments and between the state and national assessments? 39

Question 13:
What types of reports does NAEP produce? What reports are planned for the 1999 and 2000 assessments? 41

Using NAEP Data 43

Question 14:
What contextual background data does NAEP provide? 43

Question 15:
How can educators use NAEP resources such as frameworks, released questions, and reports in their work? 46

Question 16:
How are NAEP data and assessment results used to further explore education and policy issues? What technical assistance does NAEP provide? 48

Question 17:
Can NAEP results be linked to other assessment data? 50

Sampling and Data Collection 53

Question 18:

Who are the students assessed by NAEP? 53

Question 19:

How many schools and students participate in NAEP? When are the data collected during the school year? 56

Question 20:

How does NAEP use matrix sampling? What is focused BIB spiraling, and what are its advantages for NAEP? 59

Question 21:

What are NAEP's procedures for collecting data? 62

Bibliography 65

Further Reading 68

Glossary 70

Subject Index 74

INTRODUCTION

As mandated by Congress, the National Assessment of Educational Progress (NAEP) surveys the educational accomplishments of U.S. students and monitors changes in those accomplishments. NAEP tracks the educational achievements of fourth-, eighth-, and twelfth-grade students over time in selected content areas. For 30 years, NAEP has been collecting data to provide educators and policymakers with accurate and useful information.

About NAEP

Each year, NAEP employs the full-time equivalent of more than 125 people, and as many as 5,000 people work on NAEP in some capacity. These people work for many different organizations that must coordinate their efforts to conduct NAEP. Amendments to the statute that authorized NAEP established the structure for this cooperation in 1988.

Under the current structure, the Commissioner of Education Statistics, who heads the National Center for Education Statistics (NCES) in the U.S. Department of Education, is responsible, by law, for carrying out the NAEP project through competitive awards to qualified organizations. The Associate Commissioner for Assessment executes the program operations and technical quality control.

The National Assessment Governing Board (NAGB), appointed by the Secretary of Education but independent of the department, governs the program. Authorized to set policy for NAEP, the Governing Board is broadly representative of NAEP's varied audiences. NAGB selects the subject areas to be assessed, develops guidelines for reporting, and gives direction to NCES. While overseeing NAEP, NAGB often works with several other organizations. In the past, NAGB has contracted with the Council of Chief State School Officers (CCSSO) to ensure that content is planned through a national consensus process, and it contracts with ACT Inc. to identify achievement standards for each subject and grade tested.

NCES also relies on the cooperation of private companies for test development and administration services. Since 1983, NCES has conducted the assessment through a series of contracts, grants, and cooperative agreements with Educational Testing Service (ETS) and other contractors. Under these agreements, ETS is directly responsible for developing the assessment instruments, scoring student responses, analyzing the data, and reporting the results. NCES also has a cooperative agreement with Westat. Under this agreement, Westat selects the school and student samples, trains assessment administrators, and manages field operations (including assessment administration and data collection activities). National Computer Systems (NCS), which serves as a subcontractor to ETS, is responsible for printing and distributing the assessment materials and for scanning and scoring students' responses. American Institutes for Research (AIR), which serves as a subcontractor to ETS, is responsible for development of the background questionnaires.

NCES publishes the results of the NAEP assessments and releases them to the media and public. NCES strives to present this information in the most accurate and useful manner possible, publishing reports designed for the general public and specific audiences and making the data available to researchers for secondary analyses.

About the Guide

The goals of *The NAEP Guide* are to provide readers with an overview of the project and to help them better understand the philosophical approach, procedures, analyses, and psychometric underpinnings of NAEP. This guide acquaints readers with NAEP's informational resources, demonstrates how NAEP's design matches its role as an indicator of national educational achievement, and describes some of the methods used in the 1999 and 2000 assessments.

The guide follows a question-and-answer format, presenting the most commonly asked questions and following them with succinct answers. Each answer also includes additional background information. The guide is designed for the general public, including state and national policymakers; state, district, and school education officials who participate in NAEP; and researchers who rely on the guide for their introduction to NAEP.

Question: 1

What is NAEP?

Answer

Often called the "Nation's Report Card," the National Assessment of Educational Progress (NAEP) is the only nationally representative, continuing assessment of what America's students know and can do in various subject areas. NAEP provides a comprehensive measure of students' learning at critical junctures in their school experience.

The assessment has been conducted regularly since 1969. Because it makes objective information about student performance available to policymakers at national and state levels, NAEP plays an integral role in evaluating the conditions and progress of the nation's education. Under this program, only information related to academic achievement is collected, and NAEP guarantees that all data related to individual students and their families remain confidential.

FURTHER DETAILS

Overview of NAEP

Over the years, NAEP has evolved to address questions asked by policymakers, and NAEP now refers to a collection of national and state-level assessments.

Between 1969 and 1979, NAEP was an annual assessment. From 1980 through 1996, it was administered every two years. In 1997, NAEP returned to annual assessments. Initiated in 1990, state-level NAEP enables participating states to compare their results with those of the nation and other participating states.

NAEP has two major goals: to reflect current educational and assessment practices and to measure change reliably over time. To meet these dual goals, NAEP selects nationally representative samples of students who participate in either the *main* NAEP assessments or the *long-term trend* NAEP assessments.

National NAEP

National NAEP reports information for the nation and for specific geographic regions of the country (Northeast, Southeast, Central, and West). It includes students drawn from public and nonpublic schools. At the national level, NAEP is divided into two assessments: the main NAEP and the long-term trend NAEP. These assessments use distinct data collection procedures, separate samples of students, and test instruments based on different frameworks. Student and teacher background questionnaires also vary between the main and long-term trend assessments, as do many of the analyses employed to produce results. The results from these two assessments are also reported separately.

Main NAEP

The main assessments report results for grade samples (grades 4, 8, and 12). They periodically measure students' achievement in reading, mathematics, science, writing, U.S. history, civics, geography, and other subjects. (See the inside back cover.) In 2000, main NAEP will assess mathematics and science at grades 4, 8, and 12 and reading at grade 4.

The main assessments follow the curriculum frameworks developed by the National Assessment Governing Board (NAGB) and use the latest advances in assessment methodology. Indeed, NAEP has pioneered many of these innovations. The assessment instruments are flexible so they can adapt to changes in curricular and educational approaches. For example, NAEP assessments include large percentages of constructed-response questions (questions that ask students to write responses ranging from two or three sentences to a few paragraphs) and items that require the use of calculators and other materials.

As the content and nature of the NAEP instruments evolve to match instructional practices, however, the ability of the assessment to measure change over time is greatly reduced. Recent main NAEP assessment instruments have typically been kept stable for relatively short periods of time, allowing short-term trend results to be reported. For example, the 1998 reading assessment followed a *short-term* trend line that began in 1992 and continued in 1994. Because of the flexibility of the main assessment instruments, the long-term trend NAEP must be used to reliably measure change over longer periods of time.

Long-Term Trend NAEP

The long-term trend assessments report results for age/grade samples (9-year-olds/fourth grade; 13-year-olds/eighth grade; and 17-year-olds/eleventh grade). They measure students' achievements in mathematics, science, reading, and writing. Measuring trends of student achievement, or change over time, requires the precise replication of past procedures. Therefore, the long-term trend instrument does not evolve based on changes in curricula or in educational practices.

The long-term trend assessment uses instruments that were developed in the 1970s and 1980s and are administered every two years in a form identical to the original one. In fact, the assessments allow NAEP to measure trends from 1969 to the present. In 1999, the long-term trend assessment began to be administered on a four-year schedule and in different years from the main and state assessments in mathematics, science, reading, and writing.

State NAEP

Until 1990, NAEP was a *national* assessment. Because the national NAEP samples were not, and are not currently, designed to support the reporting of accurate and representative state-level results, in 1988 Congress passed legislation authorizing a voluntary Trial State Assessment (TSA). Separate representative samples of students are selected for each jurisdiction that agrees to participate in TSA, to provide these jurisdictions with reliable state-level data concerning the achievement of their students. Although the first two NAEP TSAs in 1990 and 1992 assessed only public school students, the

1994 TSA included public and nonpublic schools. Certain nonstate jurisdictions, such as U.S. territories, the District of Columbia, and Department of Defense Education Activity Schools, may also participate in state NAEP.

In 1996, "Trial" was dropped from the title of the assessment based on numerous evaluations of the TSA program. The legislation, however, still emphasizes that the state assessments are developmental.

In 1998, state NAEP assessed reading at grades 4 and 8 and writing at grade 8. In state NAEP, 44 jurisdictions participated for reading at grade 4, 41 jurisdictions for reading at grade 8, and 40 jurisdictions for writing at grade 8. In 2000, state NAEP will assess mathematics and science at grades 4 and 8.

Background Questionnaires

What factors are related to higher scores? Who is teaching students? How do schools vary in terms of courses offered? NAEP attempts to answer these questions and others through data collected on background questionnaires.

Students, teachers, and principals complete these questionnaires to provide NAEP with data about students' school backgrounds and educational activities. Students answer questions about the courses they take, homework, and home factors related to instruction. Teachers answer questions about their professional qualifications and teaching activities, while principals answer questions about school-level practices and policies. Relating student performance on the cognitive portions of the assessments to the information gathered on the background questionnaires increases the usefulness of NAEP findings and provides the context for a better understanding of student achievement.

Related Questions:

Question 14: *What contextual background data does NAEP provide?*

Question 18: *Who are the students assessed by NAEP?*

Question: 2

What subjects does NAEP assess? How are the subjects chosen, and how are the assessment questions determined? What subjects were assessed in 1999? What subjects will be assessed in 2000?

Answer

Since its inception in 1969, NAEP has assessed numerous academic subjects, including mathematics, science, reading, writing, world geography, U.S. history, civics, social studies, and the arts. (A chronological list of the assessments from 1969 to 2000 is on the inside back cover.)

Since 1988, the National Assessment Governing Board (NAGB) has selected the subjects assessed by NAEP. Furthermore, NAGB oversees creation of the frameworks that underlie the assessments and the specifications that guide the development of the assessment instruments. The framework for each subject area is determined through a consensus process that involves teachers, curriculum specialists, subject-matter specialists, school administrators, parents, and members of the general public.

In 1999, the long-term trend assessments in mathematics, science, reading, and writing were conducted using the age/grade samples described earlier (see page 4). At the national level, the 2000 assessment will include mathematics and science at grades 4, 8, and 12 and reading at grade 4. At the state level, NAEP will include mathematics and science at grades 4 and 8.

FURTHER DETAILS

Selection of Subjects

The legislation authorizing NAEP charges NAGB with determining the subjects that will be assessed. The table on page 7 identifies the subjects and grades assessed in the 1999 assessment and those in the assessment planned for 2000.

Development of Frameworks

NAGB uses an organizing framework for each subject to specify the content that will be assessed. The framework is the blueprint that guides the development of the assessment instrument.

Developing a framework can involve the following elements:

- widespread participation and reviews by educators and state education officials in the particular field of interest;
- reviews by steering committees whose members represent policymakers, practitioners, and the general public;
- involvement of subject supervisors from the education agencies of prospective participants;
- public hearings; and
- reviews by scholars in that field, by National Center for Education Statistics (NCES) staff, and by a policy advisory panel.

The *Framework* publications for the NAEP 1999 and 2000 assessments provide more details about the consensus process, which is unique for each subject.

Although they guide the development of assessment instruments, frameworks cannot encompass everything that is

taught in all the classrooms of the nation, much less everything that *should* be taught. Nevertheless, frameworks capture a range of subject-specific content and thinking skills that students need to learn and the complex issues they encounter inside and outside their classrooms. Furthermore, the consensual process used to develop the frameworks ensures that they are appropriate for current educational requirements.

Because the assessments must remain flexible to mirror changes in educational objectives and curricula, the frameworks must be forward looking and responsive, balancing current teaching practices with research findings. This flexibility is evident in the evolution of NAEP assessment instruments. The present instruments allocate a majority of testing time to constructed-response questions that require students to compose written answers. Because they require students to describe, interpret, and explain, these questions elicit a broader range of students' cognitive and analytical abilities than do simple multiple-choice questions. Furthermore, the information obtained through constructed-response questions enhances NAEP's ability to track progress toward national education goals.

Specification of Assessment Questions

Under the direction of Educational Testing Service (ETS), teachers, subject-matter specialists, and measurement experts develop the questions and tasks based on the subject-specific frameworks.

For each subject-area assessment, a national committee of experts provides guidance and reviews the questions to ensure that they meet the framework

NAEP 1999 and 2000 Assessments

1999 NAEP

Long-Term Trend	**Age 9**	**Age 13**	**Age 17**
Reading	✓	✓	✓
Mathematics	✓	✓	✓
Science	✓	✓	✓
	Grade 4	**Grade 8**	**Grade 11**
Writing	✓	✓	✓

2000 NAEP

Main NAEP	**Grade 4**	**Grade 8**	**Grade 12**
Reading	✓		
Mathematics	✓	✓	✓
Science	✓	✓	✓
State NAEP	**Grade 4**	**Grade 8**	**Grade 12**
Mathematics	✓	✓	
Science	✓	✓	

BEST COPY AVAILABLE

specifications. For each state-level assessment, the state curriculum and testing directors who comprise the NAEP NETWORK review the questions that will be included in the NAEP state component.

Framework for the 1999 Long-Term Trend Assessments

Because the long-term trend assessments measure trends in student achievement, or change over time, they must precisely replicate past procedures and frameworks. The long-term trend instruments do not evolve based on changes in curricula or in educational practices. Main NAEP assessments follow the curriculum frameworks developed by NAGB and use the latest advances in assessment methodology. As a result, the subjects of mathematics, science, reading, and writing are assessed both in long-term trend and in main NAEP, using different instruments and for different purposes.

In 1999, the long-term trend assessments were identical to the assessments begun in 1984 for reading and writing and in 1986 for mathematics and science. (Note that trend was measured several years before that because of statistical links with previous years' assessments; the chart on the inside back cover shows specific subjects and years in trend measurement.)

The reading framework for long-term trend is described in *Reading Objectives: 1983–84 Assessment*. There are four objectives for student achievement:

- Comprehending What Is Read;
- Extending Comprehension;
- Managing the Reading Experience; and
- Valuing Reading.

The writing framework for long-term trend is described in *Writing Objectives: 1984 Assessment* and *Writing Objectives: 1988 Assessment*. There are four objectives for student achievement:

- Writing to Accomplish a Variety of Purposes;
- Managing the Writing Process;
- Controlling the Forms of Written Language; and
- Valuing Writing and Written Works.

The mathematics framework for long-term trend is described in *Math Objectives: 1985–86 Assessment*. There are seven content areas:

- Fundamental Methods of Mathematics;
- Discrete Mathematics;
- Data Organization and Interpretation;
- Measurement;
- Geometry;
- Relations, Functions, and Algebraic Expressions; and
- Numbers and Operations.

The science framework for long-term trend is described in *Science Objectives: 1985–86 Assessment* and *Science Objectives: 1990 Assessment*. There are six content areas:

- Life Science;
- Physics;
- Chemistry;
- Earth and Space Science;
- History of Science; and
- Nature of Science.

The design of the assessments, sampling, and data collection are described in

the Procedural Appendix of *NAEP 1996 Trends in Academic Progress*.

Framework for the 2000 NAEP Mathematics Assessment

The framework for the 2000 NAEP mathematics assessment covers five content strands:

- Number Sense, Properties, and Operations;
- Measurement;
- Geometry and Spatial Sense;
- Data Analysis, Statistics, and Probability; and
- Algebra and Functions.

The distribution of questions among these strands is a critical feature of the assessment design, as it reflects the relative importance and value given to each of the curricular content strands within mathematics. Over the past six NAEP assessments in mathematics, the content strands have received differential emphasis. There has been continuing movement toward a more even balance among the strands and away from the earlier model, in which questions that were classified as number facts and operations accounted for more than 50 percent of the assessment item bank. Another significant difference in the newer NAEP mathematics assessments is that questions may be classified into more than one strand, underscoring the connections that exist between different mathematical topics.

A central feature of student performance that is assessed by NAEP mathematics is "mathematical power." Mathematical power is characterized as a student's overall ability to gather and use mathematical knowledge through:

- exploring, conjecturing, and reasoning logically;
- solving nonroutine problems;
- communicating about and through mathematics; and
- connecting mathematical ideas in one context with mathematical ideas in another context or with ideas from another discipline in the same or related contexts.

To assist in the collection of information about students' mathematical power, assessment questions are classified not only by content, but also by mathematical ability. The mathematical abilities of problem solving, conceptual understanding, and procedural knowledge are not separate and distinct factors of a student's ways of thinking about a mathematical situation. They are, rather, descriptions of the ways in which information is structured for instruction and the ways in which students manipulate, reason with, or communicate their mathematical ideas. As such, some questions in the assessment may be classified into more than one of these mathematical ability categories. Overall, the distribution of all questions in the mathematics assessment is approximately equal across the three categories.

Framework for the 2000 NAEP Science Assessment

The 2000 NAEP science assessment framework is organized along two major dimensions:

- Fields of science: Earth, Physical, and Life Sciences; and
- Knowing and doing science: Conceptual understanding, Scientific investigation, and Practical reasoning.

Fields of science. The assessment emphasizes knowledge in the content areas. The content assessed in earth science centers on objects and events that are relatively assessable or visible, such as Earth (lithosphere), water (hydrosphere), air (atmosphere), and the Earth in space. The physical science component relates to basic knowledge and understanding concerning the structure of the universe, as well as the physical principles that operate within it. The assessment probes matter and its transformations, energy and its transformations, and the motion of things. Major concepts assessed in life sciences include change and evolution, cells and their functions, organisms, and ecology.

Knowing and doing science. This dimension stresses the connections and organization of factual knowledge in science. Conceptual understanding is the measure of students' abilities to perceive and grasp the meaning of ideas. Students should acquire a rich collection of scientific information that will enable them to move from simply being able to provide reasonable interpretations of observations to providing explanations and predictions. Assessment exercises in grade 4 deal with students' ability to elaborate principles from personal experiences. In grades 8 and 12, the emphasis shifts from richness of experience to reasonable scientific interpretation of observations. This database of organized information also allows students to apply information efficiently in the design and execution of scientific investigations and in practical reasoning.

Scientific investigation represents the activities of science that distinguish it from other ways of knowing about the world. Science involves designing fair tests and considering all the variables and the means to control the variables. As students are asked to demonstrate their ability to do scientific investigations, it is important to keep in mind their development in understanding and performance, not just with respect to the control of variables but also regarding the other elements of science. Appropriate to their age and grade level, students will be assessed on their ability to acquire new information, plan appropriate investigations, use a variety of scientific tools, and communicate the results of their investigations.

Practical reasoning is a complex skill that develops throughout life. As children mature they also learn to take a depersonalized view of a situation and to consider someone else's point of view. Practical reasoning should become a major factor in science assessment at grades 8 and 12 rather than at grade 4. By grade 12, students should be able to discuss larger science- and technology-linked problems not directly related to their immediate experience. Examples include waste disposal, energy uses, air quality, water pollution, noise abatement, and the tradeoffs between the benefits and adverse consequences of various technologies. Practical reasoning includes competence in analyzing a problem, planning appropriate approaches, evaluating them, carrying out the required procedures for the approach selected, and evaluating the results.

In addition to the two major dimensions, the framework includes two other categories that pertain to a limited subset of items:

- Nature of science; and
- Themes.

Nature of science includes the historical development of science and technology, the habits of mind that characterize these fields, and methods of inquiry and prob-

lem solving. Themes represent big ideas or key organizing concepts that pervade science. Themes include the ideas of *systems* and their application in the disciplines, *models* and their function in the development of scientific understanding and its application to practical problems, and *patterns of change* as exemplified in natural phenomena.

Framework for the 2000 NAEP Reading Assessment

The NAEP reading assessment framework, used from 1992 to 2000 and grounded in current theory, views reading as a dynamic, complex interaction that involves the reader, the text, and the context of the reading experience. As specified in the framework, the assessment addresses three purposes for reading:

- reading for literary experience;
- reading for information; and
- reading to perform a task.

Reading for literary experience involves reading novels, short stories, poems, plays, and essays to learn how authors present experiences and interaction among events, emotions, and possibilities. Reading to be informed involves reading newspapers, magazine articles, textbooks, encyclopedias, and catalogues to acquire information. Reading to perform a task involves reading documents such as bus schedules, directions for a game, laboratory procedures, recipes, or maps to find specific information, understand the information, and apply it. (Reading to perform a task is not assessed at grade 4.)

Within these purposes for reading, the framework recognizes four ways that readers interact with text to construct meaning from it. These four modes of interaction, called "reading stances," are as follows:

- forming an initial understanding;
- developing an interpretation;
- engaging in personal reflection and response; and
- demonstrating a critical stance.

All reading assessment questions are developed to reflect one of the purposes for reading and one of the reading stances.

The following questions from a previous grade 4 reading assessment indicate the reading purposes and stances tested by the questions and illustrate a sample student response.

Grade 4

Story:

Hungry Spider and the Turtle

"Hungry Spider and the Turtle" is a West African folktale that humorously depicts hunger and hospitality through the actions and conversations of two very distinct characters. The ravenous and generous Turtle, who is tricked out of a meal by the gluttonous and greedy Spider, finds a way to turn the tables and teach the Spider a lesson.

Questions:

Why did Spider invite Turtle to share his food?

A. To amuse himself

B. To be kind and helpful

C. To have company at dinner

▸ D. To appear generous

Reading Purpose: Literary Experience

Question 2

Reading Stance: Developing an Interpretation

Who do you think would make a better friend, Spider or Turtle? Explain why.

Reading Purpose: Literary Experience

Reading Stance: Personal Response

Sample Response:

I think Turtle because insted of get angry with Spider he just pad him a lesson.

For further discussion of the framework, see the *Reading Framework for the National Assessment of Educational Progress: 1992–2000*. For additional explanations and the results, see *NAEP 1998 Reading Report Card for the Nation and the States*.

Related Questions:

Question 1: *What is NAEP?*

Question 6: *What process is used to develop the assessments?*

The NAEP 2000 Reading Assessment
Aspects of Reading Literacy

Constructing, Extending, and Examining Meaning

	Initial Understanding	Developing Interpretation	Personal Reflection and Response	Demonstrating a Critical Stance
	Requires the reader to provide an initial impression or unreflected understanding of what was read.	Requires the reader to go beyond the initial impression to develop a more complete understanding of what was read.	Requires the reader to connect knowledge from the text to his or her own personal background. The focus is on how the text relates to personal knowledge.	Requires the reader to stand apart from the text and consider it.
Reading for Literary Experience	What is the story/plot about?	How did the plot develop?	How did this character change your idea of ___?	Rewrite this story with ___ as a setting or ___ as a character.
	How would you describe the main character?	How did this character change from the beginning to the end of the story?	Is this story similar to or different from your own experiences?	How does this author's use of ___ (irony, personification, humor) contribute to ___?
Reading for Information	What does this article tell you about ___?	What caused this event?	What current event does this remind you of?	How useful would this article be for ___? Explain.
	What does the author think about this topic?	In what ways are these ideas important to the topic or theme?	Does this description fit what you know about ___? Why?	What could be added to improve the author's argument?
Reading to Perform a Task	What is this supposed to help you do?	What will be the result of this step in the directions?	In order to ___, what information would you need to find that you don't know right now?	Why is the information needed?
	What time can you get a nonstop flight to X? (Search)	What must you do before this step?	Describe a situation where you could leave out step X.	What would happen if you omitted this?

BEST COPY AVAILABLE

Question: 3

Is participation in NAEP voluntary? Are the data confidential? Are students' names or other identifiers available?

Answer

Federal law specifies that NAEP is *voluntary* for every pupil, school, school district, and state. Even if selected, school districts, schools, and students can refuse to participate without facing any adverse consequences from the federal government. Some state legislatures mandate participation in NAEP, others leave the option to participate to their superintendents and other education officials at the local level, and still other states choose not to participate.

Federal law also dictates that NAEP data remain *confidential*. The legislation authorizing NAEP—the National Education Statistics Act of 1994, Title IV of Improving America's Schools Act of 1994, U.S.C. 9010—stipulates in Section 411(c)(2)(A):

> The Commissioner shall ensure that all personally identifiable information about students, their education performance, and their families, and that information with respect to individual schools, remains confidential, in accordance with Section 552a of Title 5, United States Code.

After publishing NAEP reports, the National Center for Education Statistics (NCES) makes the data available to researchers but withholds students' names and other identifying information. Although it might be possible for researchers to deduce the identities of some NAEP schools, they must swear to keep these identities confidential, under penalty of fines and jail terms, before gaining access to NAEP data.

FURTHER DETAILS

A Voluntary Assessment

Participation in NAEP is voluntary for states, school districts, schools, teachers, and students. Participation involves responding to test questions that focus on a particular subject and to background questions that concern the subject area, classroom practices, school characteristics, and student demographics. Answering any of these questions is voluntary.

Before any student selected to participate in NAEP actually takes the test, the student's parents decide whether or not their child will do so. Local schools determine the procedures for obtaining parental consent.

NAEP background questions provide educators and policymakers with useful information about the educational environment. Nonparticipation and nonresponse—by students as well as teachers—greatly reduce the amount of potentially helpful information that can be reported.

A Confidential Assessment

All government and contractor employees who work with NAEP data swear to uphold a confidentiality law. If any employee violates the confidentiality law by disclosing the identities of NAEP respondents, that person is subject to criminal

penalties, including fines and prison terms.

During test administration, the names of students are used to assign specific test booklets to students selected for a particular assessment. Each booklet has a unique identification number so that it can be linked to teacher and school data. After the booklets have been completed and absent students have taken makeup tests, NAEP no longer needs students' names, and the links between students' names and their test booklets are destroyed.

NAEP administrators use tear-off forms to break the link between the names and identification numbers. When the booklets are sent to NAEP for scoring, the portion of the form containing students' names remains in the school. School officials keep these forms in a secure storage envelope for a few weeks after the assessment in case the link to the identification numbers needs to be checked. When the information is no longer needed, schools are notified and officials destroy the storage envelope, confirming their actions by returning a Destruction Notice to NAEP.

Released Data

Most people make use of published reports and other NAEP summary data such as almanacs. However, educational researchers may have an interest in additional analyses that require access to raw NAEP data. Because public funds are used for NAEP, these raw data are made available after their collection to researchers through restricted-use data tapes, subject to approval by NCES. The released data do not include names, addresses, or other personally identifiable information about the students who were assessed. In those extremely rare cases in which schools or teachers might be identified—because the data they reported were unusual or unique—NCES suppresses sufficient information to eliminate that possibility.

To receive these restricted-use data files, NCES requires researchers to submit sworn statements vowing not to disclose any identifiable information. These statements must be provided before researchers access the data. Researchers who violate the confidentiality law are subject to the same criminal penalties—fines and prison terms—as government and contractor employees.

Related Question:

Question 4: *Can parents examine the questions NAEP uses to assess student achievement? Can parents find out how well their children performed in the NAEP assessment? Why are NAEP questions kept confidential?*

Question: 4

Can parents examine the questions NAEP uses to assess student achievement? Can parents find out how well their children performed in the NAEP assessment? Why are NAEP questions kept confidential?

Answer

Every parent has the right of access to the educational and measurement materials that their children encounter. NAEP provides a demonstration booklet so that interested parents may review questions similar to those in the assessment. Under certain prearranged conditions, small groups of parents can review the booklets being used in the actual assessment. This review must be arranged with the school principal, NAEP field supervisor, or school coordinator, who will ensure that test security is maintained.

NAEP is not designed, however, to report scores for individual students. So, although parents may examine the NAEP test questions, the assessment yields no scores for their individual children.

As with other school tests or assessments, most of the questions used in NAEP assessments remain secure or confidential to protect the integrity of the assessment. NAEP's integrity must be protected because certain questions measure student achievement over a period of time and must be administered to students who have never seen them before.

Despite these concerns, NAEP releases nearly one-third of the questions used in each assessment, making them available for public use. Furthermore, the demonstration booklets provided by NAEP make all student background questions readily available for review.

FURTHER DETAILS

Parent Access to NAEP Booklets

Because parents are interested in their children's experiences in school, NAEP provides the school with a demonstration booklet before the assessment is scheduled. This demonstration booklet, which may be reproduced, contains all student background questions and sample cognitive questions. Parents can obtain copies of the demonstration booklet from the school.

Within the limits of staff and resources, school administrators and parents can review the NAEP booklets being used for the current assessment. Arrangements for this review must be made prior to the local administration dates so that sufficient materials can be prepared and interested persons can be notified of its time and location. Upon request, NAEP staff will also review the booklets with small groups of parents, with the understanding that no assessment questions will be duplicated, copied, or removed.

Requests for these reviews can be made to the NAEP data collection staff or by contacting the National

Center for Education Statistics (NCES) at 202–219–1831. Individuals whose children are not participating in the assessment but who wish to examine secure assessment questions can contact the U.S. Department of Education's Freedom of Information Act officer at 202–708–4753.

The Importance of Security

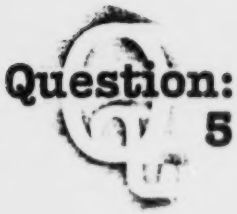
Measuring student achievement and comparing students' scores from previous years requires reusing some questions for continuity and statistical purposes. These questions must remain secure to assess trends in academic performance accurately and to report student performance on existing NAEP score scales.

Furthermore, for NAEP to regularly assess what the nation's students know and can do, it must keep the assessment from being compromised. If students have prior knowledge of test questions, then schools and parents will not know whether their performance is based on classroom learning or coaching on specific assessment questions. After every assessment, nearly one-third of the questions are released to the public. These questions can be used for teaching or research. NAEP reports often contain samples of actual questions used in the assessments. Sample questions can also be obtained from NCES, NAEP Released Exercises, 555 New Jersey Avenue, NW, Washington, DC 20208–5653 or on the Web site at http://nces.ed.gov/nationsreportcard.

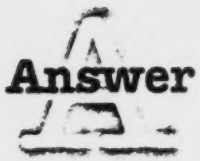
Related Questions:

Question 3: *Is participation in NAEP voluntary? Are the data confidential? Are students' names or other identifiers available?*

Question 15: *How can educators use NAEP resources such as frameworks, released questions, and reports in their work?*



Who evaluates and validates NAEP?



NAEP and its findings have a considerable impact on the public's understanding of student academic achievement. Because NAEP plays a unique and prominent role, precautions must be taken to ensure the validity and reliability of its findings. Therefore, Congress consistently passes legislation that establishes panels to evaluate the assessment as a whole. In response to these mandates, the National Center for Education Statistics (NCES) has established various expert panels to study NAEP. These panels have produced a series of reports that address numerous important NAEP issues.

FURTHER DETAILS

The Technical Review Panel

By law, the Commissioner of NCES must provide "continuing reviews of the National Assessment, including validation studies" (P.L. 100–297, Sec. 3403 [I] [9] [A]). In fulfillment of this mandate, the Center for Research on Evaluation, Standards, and Student Testing (CRESST) at the University of California, Los Angeles, in conjunction with the University of Colorado at Boulder and RAND, was awarded a contract in 1989 to establish the Technical Review Panel (TRP).

Beginning in 1989, TRP produced a series of studies on specific questions about the validity of interpreting NAEP results, including:

- the quality of NAEP data;
- the number and character of NAEP scales;
- the robustness of NAEP trend lines;
- the trustworthiness of one interpretation of group comparisons;
- the validity of interpretations of NAEP anchor points and achievement levels;
- the linking of other test results to NAEP;
- the effects of student motivation on performance;
- the adequacy of NAEP data about student background and instructional experiences; and
- the understanding educators and policymakers obtain from NAEP reports.

At the end of the project period, TRP produced a final report summarizing the results of its studies (Linn, Koretz, & Baker, 1996).

The Trial State Assessment

Another panel was commissioned to study the validity of the NAEP state component. In spring 1988, Congress enacted Public Law 100–297, which authorized

the NAEP Trial State Assessment (TSA). In authorizing the TSA, Congress called for an independent evaluation of "the feasibility and validity of state assessments and the fairness and accuracy of the data they produce." In response to the legislation, NCES awarded a grant in 1990 to the National Academy of Education (NAE) Panel on the Evaluation of the NAEP Trial State Assessment Project.

Between 1992 and 1996, the NAE Panel produced numerous reports that evaluated the validity of TSA. The panel released a capstone report in April 1997. This final report concluded that in less than 10 years NAEP had increased by a factor of four the number of students assessed; undergone important changes in test content, design, and administration; and engaged the intense interest of many stakeholders and observers. The rapid pace of these changes created extreme conditions of conflicting demands, strained resources, and technical complexities that were a potential threat to the existence of the entire program.

Six major developments have become points of continuous discussion in NAEP: expansion of the assessment to include state-level NAEP; inclusion of more challenging performance tasks; testing broader and more representative samples of students, including students with disabilities or limited English proficiency; pressure to make NAEP standards-based in the absence of nationally agreed-upon content and performance standards; desire for international comparisons; and desire to link NAEP with state assessments. With the importance of assessment in education reform likely to continue and increase, NAEP's future course depends on critical decisions that take all these factors into consideration. The NAE Panel also evaluated the National Assessment Governing Board performance standards in a separate report.

Evaluation

In 1996, the National Academy of Sciences (NAS) was awarded a contract to further evaluate the national and state NAEP. In response, NAS formed a committee of distinguished educators and other experts to conduct the evaluation activities described in the Congressional mandate of 1994. Public Law 103–382 mandates that "the Secretary shall provide for continuing review of the National Assessment, State Assessments, and student performance levels by one or more nationally recognized organizations." In the evaluation process, the committee directed workshops; commissioned papers; solicited testimony and interviews; observed NAEP activities; and studied program documents, extant research, and prior evaluation reports. NAS released its NAEP evaluation report, *Grading the nation's report card: Evaluating NAEP and transforming the assessment of educational progress*, in early 1999. The report presented observations and recommendations in a number of key areas, including streamlining the design of NAEP, enhancing the participation and meaningful assessment of English-language learners and students with disabilities, framework design and the assessment development process, and the setting of reasonable and useful performance standards. NCES has requested that NAS continue to evaluate specific aspects of the NAEP program in the coming years.

A new committee is looking at issues surrounding district-level reporting and market basket reporting, two important issues raised in the NAS evaluation report.

The NAEP Validity Studies Panel

NCES funded a contract that established the NAEP Validity Studies (NVS) Panel. This panel was formed to provide technical review of NAEP plans and products and to identify technical concerns and promising techniques worthy of further study and research.

Since its inception in October 1995, the NVS Panel has worked on numerous validity studies, and that work will continue in the coming years. To date, the panel has released reports on: (1) optimizing state NAEP, (2) the information value of constructed-response items, (3) evaluation of NAEP equating procedures, (4) a feasibility study of two-stage testing in large-scale educational assessments, (5) the effects of finite sample corrections on state assessment sample size requirements, (6) a proposed research program on NAEP reporting, and (7) an investigation of why some students do not respond to NAEP questions.

Question: 6

What process is used to develop the assessments?

Answer

To meet the nation's growing need for information about what students know and can do, the NAEP assessment instruments must measure change over time and must reflect changes in curricula and instruction in diverse subject areas. Meeting these goals can be especially challenging because instructional design and objectives may change at any time in the nation's 100,000 schools.

Developing the assessment instruments—from writing questions to analyzing field-test results to constructing the final instruments—is a complex process that consumes most of the time during the interval between assessments. In addition to conducting a field test, developers subject the assessment instruments to numerous reviews to identify areas that require revision or augmentation so they comply with the specifications of the framework and the achievement levels.

FURTHER DETAILS

The Development Process

The following section summarizes the instrument development process that NAEP uses for the main and state assessments. Newly developed assessment questions and exercises go through all the steps in this process. The many reviews help identify areas that must be revised or augmented to ensure compliance with the framework and the achievement levels. Thus, many experts offer input into the development process, ensuring that the tests adhere as closely as possible to the goals established by the National Assessment Governing Board (NAGB).

Summary of the NAEP Instrument Development Process

- Educational Testing Service (ETS) test development specialists and various subject-matter experts write the questions and exercises and then classify them according to framework specifications.
- Test development staff experienced in the subject area review the questions and exercises for content concerns and revise them accordingly.
- Questions and exercises are banked in the test development system, as is all classification information.
- A test developer assembles blocks of questions and exercises for field tests according to specifications.
- Specialists review the blocks, which undergo mandatory sensitivity and editorial reviews.
- Assessment questions are administered to individual students in one-on-one question tryout sessions to determine both how well students understand the questions and what further refinements should be made

to the wording or formatting of questions.

- Instrument Development Committees convene to review the questions and blocks and to independently confirm that the questions fit the framework specifications and are correctly classified.
- Outside groups of content and assessment experts verify the question classifications independently.
- For the state assessment program, the NAEP NETWORK reviews all questions, exercises, blocks, and questionnaires that will be included in the assessment.
- A test developer updates the test development version of the questions and exercises based on committee, NAEP NETWORK, and content and assessment expert reviews.
- The field test questionnaires and assessment exercises are reviewed by the National Center for Education Statistics (NCES), the National Assessment Governing Board (NAGB), the Office of Management and Budget (OMB), and the Information Management Compliance Division (IMCD) for compliance with government policies on data collection. Revisions to the field test versions are made as needed to obtain government clearance.
- A clearance number is obtained for the field test.
- Field test booklets and questionnaires are printed, and other assessment materials (such as audiotapes, photographs, and hands-on science materials) are produced.
- Exercises from the blocks are banked in the test development system.
- **The field tests are administered.**
- The field tests are scored and analyzed.
- Suitable questions for the assessment are selected.
- Subject-matter specialists review the blocks selected for the assessment.
- The blocks undergo sensitivity and editorial reviews.
- Instrument Development Committees convene to review the questions and blocks and to independently confirm classification codes.
- The assessment questionnaires and assessment exercises are reviewed by NCES, NAGB, OMB, and IMCD for compliance with government policies on data collection. Revisions to the field test versions are made as needed to obtain government clearance.
- The camera-ready blocks are proofed and approved for printing.
- Final versions of other assessment materials (such as audiotape and photographs) are approved for production.
- The assessment booklets and questionnaires are prepared, proofed, and printed.
- Exercises from the blocks are banked in the test development system.
- **The assessments are administered.**

The blocks used in NAEP are collections of questions administered to students as a timed unit. NAGB is responsible for ensuring that all questions selected for NAEP are free from racial, cultural, gender, or regional bias. Thus, the blocks undergo a

mandatory sensitivity review to ensure that the assessment reflects a thoughtful, balanced consideration of all groups of people. External reviewers, including state education agency personnel, review the questions for appropriateness for students from a variety of backgrounds and across regions. NCES also reviews all NAEP questions, and OMB and IMCD further review the background questions.

As a final quality control measure to monitor against bias, the results for each question are checked empirically after the field test. This empirical check for fairness employs differential item functioning (DIF) analyses. DIF analyses identify questions that are differentially difficult for particular subgroups of students (who are categorized by racial or ethnic group membership or by gender) for reasons that seem unrelated to the overall ability of the students. For further discussion of DIF procedures, see the *NAEP 1996 Technical Report*.

The Instrument Development Committee reviews the questions and blocks and independently confirms the classification codes. This committee meets several times during the development cycle to consider question format and appropriateness and the cognitive processes being measured, to refine scoring rubrics after the field test, and to review field test results.

The NAEP NETWORK includes representatives from nonpublic schools and assessment directors from all 50 states, Guam, Puerto Rico, the Virgin Islands, the District of Columbia, and the Department of Defense Education Activity schools. The NAEP NETWORK convenes to review all exercises, blocks, and questionnaires that will be included in the state assessment program.

Related Question:

Question 11: *How does NAEP analyze the assessment results?*

BEST COPY AVAILABLE

Question: 7

How does NAEP accommodate students with disabilities and students with limited English proficiency?

Answer

Throughout its history, NAEP has always encouraged the inclusion of all students who could meaningfully participate in the assessment, including those with disabilities and those with limited English proficiency. An estimated 10 percent of the school population is classified as having a disability or limited English proficiency. Nearly half of these students have been included in previous assessments, although the percentages vary by grade and subject being assessed.

Previously, because of concerns about standardized administration, accommodations such as bilingual booklets and extended testing time were not permitted, excluding some students who could have participated if accommodations had been made. Because it is committed to increased inclusion, the National Center for Education Statistics (NCES) formally tested new policies with the 1996 NAEP assessment. Under these guidelines, school administrators were encouraged even more than in the past to include students with disabilities or limited English proficiency (SD/LEP) if any doubt about excluding the student existed. Although NAEP establishes the criteria for inclusion, differences remain among states in how SD/LEP students are treated. Because of the 1997 amendments to the Individuals with Disabilities Education Act, some states are changing their procedures for students with disabilities.

FURTHER DETAILS

Increased Inclusion

NAEP intends to assess all students selected to participate. However, some students may have difficulty with the assessment as it is normally administered because of a disability or limited English proficiency.

Beginning with the 1996 national assessment, NAEP implemented a two-part modification of procedures to increase inclusion in NAEP assessments. First, revised criteria were developed to define how decisions about inclusion should be made. Second, NAEP provided certain accommodations that were either specified in a student's Individualized Education Plan (IEP) or frequently used to test the student. The accommodations vary depending on the subjects being assessed.

When a school identifies a student as having a disability or limited English proficiency, the teacher or staff member who is most familiar with the student is asked to complete a questionnaire about the services received by the student and to determine whether the student should take part in the assessment. The questionnaire provides useful information about differential exclusion rates across disability conditions and across the states. Students who cannot take part, even with an accommodation allowed by NAEP, are excluded from the assessment.

Summary

NAEP has traditionally included more than 90 percent of the students selected for the sample. Even though the percentage of exclusion is now relatively small, NAEP continually explores ways to further reduce exclusion rates while ensuring that NAEP results are representative and can be generalized.

Related Question:

Question 14: *What contextual background data does NAEP provide?*

Question: 8

What assessment innovations has NAEP developed?

Answer

NAEP frameworks are updated to reflect changes both in curricula and in the way subjects are taught. The 1997 NAEP arts assessment emphasized performance and integrated performance tasks into the assessment instruments for visual arts, music, dance, and theatre. The 1998 assessments in reading, writing, and civics emphasized activities that used a variety of stimulus materials such as cartoons, color photographs, and letters. Both the mathematics and science assessments for 2000 take into account recent developments in education. The mathematics assessment incorporates the use of calculators, rulers, protractors, and manipulatives into meaningful assessment exercises. Many questions in the science assessment are performance exercises, and some hands-on tasks require students to engage in scientific investigations. Since 1996, all NAEP assessments have emphasized the inclusion of more students who require special accommodations.

FURTHER DETAILS

Innovations in Recent Assessments

Each new framework contains assessment objectives that have been updated to reflect changes in curricula and instruction, often requiring innovations in assessment instrumentation, scoring procedures, and analysis methodology. Even when the same framework guides assessments for several years, shifts in curricular or instructional practice may necessitate the field testing or use of new blocks of questions or performance tasks for subsequent assessments.

NAEP in 2000

The 2000 NAEP assessment in mathematics is the fourth in a trend line that began in 1990, continuing in 1992 and in 1996. The 2000 assessment maintains the design of the earlier mathematics assessments. It also continues the trend of incorporating the use of calculators, rulers, protractors, and manipulatives into meaningful assessment exercises that are presented in both multiple-choice and constructed-response formats. Some constructed-response questions require students to generate extended responses that support conjectures, justify conclusions, and substantiate numerical results.

The 2000 NAEP assessment in science is the second in a trend line that began in 1996. Many questions in the assessment are performance exercises. These enhance the assessment's measurement of students' abilities to reason, explain, and apply scientific knowledge and to plan and evaluate scientific investigations. In addition, many students are required to complete a hands-on task that requires them to manipulate equipment, observe, measure, and reach some conclusions regarding their investigation.

The Impact of NAEP Innovations

The NAEP assessments have had an impact throughout the nation as NAEP data, the frameworks, and their objectives have served as models for several state testing programs. For example, Maryland used the 1990 NAEP reading and writing frameworks to develop the Maryland Learning Outcomes in those areas. As a result, the Maryland outcomes resemble the NAEP objectives in name and construct definition. West Virginia, like Maryland, has used NAEP as a model for its curriculum specifications.

In 1994, North Carolina conducted a special study of eighth-grade students to link their performance on the North Carolina End-of-Grade Tests in Mathematics to the NAEP mathematics assessment performance at grade 8. The state is also examining ways to better align the achievement levels from the end-of-grade tests for grades 3 through 8 with those from the NAEP assessment.

Finally, Ohio used the NAEP frameworks as models to develop its content standards. Ohio has also modeled questions on the NAEP item structure when developing state assessments at grades 4 and 6 in reading, writing, mathematics, citizenship, and science. For additional information about NAEP activities in a given state, contact the NAEP coordinator for that state. A list of coordinators appears on the NCES World Wide Web site (http://nces.ed.gov/nationsreportcard).

Question: 9

What results does NAEP provide?

Answer

NAEP provides results about subject-matter achievement, instructional experiences, and school environment and reports these results by populations of students (e.g., fourth graders) and subgroups of those populations (e.g., male students or Hispanic students). NAEP does *not* provide individual scores for the students or schools assessed.

Subject-matter achievement is reported in two ways—scale scores and achievement levels—so that student performance can be more easily understood. NAEP scale score results provide information about the distribution of student achievement by groups and subgroups. Achievement levels categorize student achievement as *Basic*, *Proficient*, and *Advanced*, using ranges of performance established for each grade. (A fourth level, below *Basic*, is also reported for this scale.) Achievement levels are used to report results by a set of standards for what students should know and be able to do.

Because NAEP scales are developed independently for each subject, scale score and achievement level results cannot be compared across subjects. However, these reporting metrics greatly facilitate performance comparisons within a subject from year to year and from one group of students to another in the same grade.

FURTHER DETAILS

NAEP Contextual Variables

As the Nation's Report Card, national NAEP examines the collective performance of U.S. students. State NAEP provides similar information for participating jurisdictions. Although it does not report on the performance of individual students, NAEP reports on the overall performance of aggregates of students (e.g., the average reading scale score for eighth-grade students or the percentage of eighth-grade students performing at or above the *Proficient* level in reading). NAEP also reports on major subgroups of the student population categorized by demographic factors such as race or ethnicity, gender, highest level of parental education, location of the school (central city, urban fringe or large town, or rural or small town), and type of school (public or nonpublic).

Information provided through background questionnaires completed by students, teachers, and school administrators enables NAEP to examine student performance in the context of various education-related factors. For instance, the NAEP 1998 assessments reported results gathered from these questionnaires for the following contextual variables: course taking, homework, use of textbooks or other instructional materials, home discussions of school work, and television-viewing habits.

NAEP Subject-Matter Achievement

NAEP assessments provide a great deal of information about students' knowledge and abilities. The usefulness of the information obtained through the main and state NAEP is maximized by presenting average scores and percentiles on NAEP subject scales and by presenting the percentages of students attaining specific NAEP achievement levels. (Results from the long-term trend component of the national assessment are reported only in terms of NAEP subject-area scale scores, not achievement levels.)

Starting with the 1990 assessment, the National Assessment Governing Board (NAGB) developed achievement levels for each subject at each grade level to measure how well students' *actual* achievement matches the achievement *desired* of them. Thus, NAEP results provide information about what students *know* and *can do* (mapped on the NAEP subject scale), and they indicate the extent to which student achievement meets expectations of what students *should know* and *should be able to do* (mapped as an achievement level).

NAEP Subject Scales

For each subject area assessed, student responses are analyzed to determine the percentage of students responding correctly to each multiple-choice question and the percentage of students performing in each of the score categories for constructed-response questions. Item response theory (IRT) methods are used to generate scales that summarize performance on the primary dimensions of the curricular frameworks used to develop the assessment. For example, three scales were developed for the NAEP 1998 reading assessments: reading for literary experience, reading to gain information, and reading to perform a task. A reading composite scale based on the weighted average of the three scales was also developed and used as the principal measure for reporting NAEP results. Because the scales for each NAEP subject are developed independently, results cannot be compared across subjects.

Achievement Level Policy Definitions

Basic	Partial mastery of prerequisite knowledge and skills that are fundamental for proficient work at each grade.
Proficient	Solid academic performance for each grade assessed. Students reaching this level have demonstrated competency over challenging subject matter, including subject-matter knowledge, application of such knowledge to real-world situations, and analytical skills appropriate to the subject matter.
Advanced	Superior performance.

Achievement Levels

For the three subjects to be assessed in 2000, results will be reported using the achievement levels authorized by the NAEP legislation. A broadly representative panel of teachers, education specialists, and members of the general public defined these achievement levels, which were adopted by NAGB. The achievement levels are based on collective judgments about what students *should know* and *should be able to do* relative to the content reflected in the NAEP assessment framework.

For each grade assessed, three achievement levels are defined: *Basic*, *Proficient*, and *Advanced*. The table on page 29 presents the policy definitions of the three achievement levels, and the *Report Card* for each subject contains detailed descriptions of the subject-specific achievement levels.

The NAEP legislation requires that the achievement levels be used on a developmental basis until the Commissioner of Education Statistics determines, as the result of a congressionally mandated evaluation by one or more nationally recognized evaluation organizations, that the achievement levels are "reasonable, valid, and informative to the public." Upon review of the available information, the Commissioner of Education Statistics agrees with the National Academy's recommendation that caution needs to be exercised in the use of the current achievement levels, since in the opinion of the Academy, "... appropriate validity evidence for the cut scores is lacking; and the process has produced unreasonable results." Therefore, the Commissioner concludes that these achievement levels should continue to be considered developmental and should continue to be interpreted and used with caution. The Commissioner and the Governing Board believe that the achievement levels are useful for reporting on trends in the educational achievement of students in the United States.

Limitations of NAEP

Simple or causal inferences related to subgroup membership, the effectiveness of public and nonpublic schools, and state- or district-level educational systems cannot be drawn using NAEP results. For example, performance differences observed among racial or ethnic subgroups are almost certainly associated with a broad range of socioeconomic and educational factors that are not addressed by NAEP assessments. In addition, the student participation rates and motivation of students, particularly twelfth graders, should be considered when interpreting NAEP results.

NAEP does not, nor is it designed to, report scores for individuals. Therefore, student-level inferences should not be drawn from the NAEP data.

The NAEP assessment results are most useful when they are considered in light of other knowledge about the education system, such as trends in educational reform, changes in the school-age population, and societal demands and expectations.

Related Questions:

Question 2: *What subjects does NAEP assess? How are the subjects chosen, and how are the assessment questions determined? What subjects were assessed in 1999? What subjects will be assessed in 2000?*

Question 14: *What contextual background data does NAEP provide?*

Question 18: *Who are the students assessed by NAEP?*

Question: 10

How does NAEP reliably score and process millions of student-composed responses?

Answer

Scoring a large number of constructed responses with a high level of reliability and within a limited time frame is essential to NAEP's success. (In 1998, approximately 3.8 million constructed responses were scored.) To ensure reliable, quick scoring, Educational Testing Service (ETS) and National Computer Systems (NCS) take the following steps:

- develop focused, explicit scoring guides that match the criteria emphasized in the assessment frameworks;
- recruit qualified and experienced scorers, train them, and verify their abilities through qualifying tests;
- employ an image-processing and scoring system that routes student responses directly to the scorers so they can focus on scoring rather than paper routing;
- monitor scorer consistency through ongoing reliability checks and assess the quality of scorer decision making through frequent backreading; and
- document all training, scoring, and quality control procedures in the technical reports.

The 2000 NAEP assessments will contain both constructed-response and multiple-choice questions. The constructed responses are scored using the image-processing system, whereas the responses to the multiple-choice questions are scored by scanning the test booklets. The following table summarizes the scoring from the 1998 assessments.

1998 NAEP Assessment: Main and State Constructed Responses Scored in Reading, Writing, and Civics

	Reading	Writing	Civics	Total
Number of Questions	191	66	78	335
Responses Scored	3,201,855	373,280	195,817	3,770,952
Elapsed Scoring Time		14 1/2 weeks		
Number of Scorers and Leaders Required		637		

BEST COPY AVAILABLE

FURTHER DETAILS

Developing Scoring Guides

Scoring guides for the assessments are developed using a multistage process.

First, scoring criteria are articulated. While the constructed-response tasks are being developed, an initial version of the scoring guides is drafted. Subject area and measurement specialists, the Instrument Development Committees, the National Center for Education Statistics (NCES), and the National Assessment Governing Board (NAGB) review the scoring guides to ensure that they include criteria consistent with the wording of the questions; are concise, explicit, and clear; and reflect the assessment framework criteria.

Next, the guides are used to score student responses from the field test. The committees and ETS staff use the results from this field test to further refine the guides. Finally, training materials are prepared. Assessment specialists from ETS select examples of student responses from the actual assessment for each performance level specified in the guides. Selecting the examples provides a final opportunity to refine the wording in the scoring guides, develop additional training materials, and make certain that the guides accurately represent the assessment framework criteria.

The examples clearly express the committees' interpretations of each performance level described in the scoring guides and help illustrate the full range of achievement under consideration. During the actual scoring process, the examples help scorers interpret the scoring guides consistently, thereby ensuring the accurate and reliable scoring of diverse responses.

Recruiting and Training Scorers

Recruiting highly qualified scorers to evaluate students' responses is crucial to the success of the assessment. A five-stage model is used for selecting and training scorers.

The first stage involves selecting scorers who meet qualifications specific to the subject areas being scored. Prospective scorers participate in a simulated scoring exercise and a series of interviews before being hired. (Some applicants take an additional exam for writing mechanics.)

Next, scorers are oriented to the project and trained to use the image scoring system. This orientation includes an in-depth presentation of the goals of NAEP and the frameworks for the assessments.

At the third stage, training materials, including sample papers, are prepared for the scorers. ETS trainers and NCS scoring supervisors read hundreds of student responses to select papers that represent the range of scores in the scoring guides while ensuring that a range of participating schools; racial, ethnic, and gender groups; geographic regions; and communities is represented in the training papers.

In the fourth stage, ETS and NCS subject-area specialists train scorers using the following procedures:

- presenting and discussing the task to be scored and the task rationale;
- presenting the scoring guide and the anchor responses;
- discussing the rationale behind the scoring guide, with a focus on the

criteria that distinguish the various levels of the guide;

- practicing the scoring of a common set of sample student responses;
- discussing in groups each response contained in the practice scoring; and
- continuing the practice steps until scorers reach a common understanding of how to apply the scoring guide to student responses.

In the final stage, scorers assigned to questions that require long constructed responses work through a qualification round to ensure that they can reliably score student responses for extended constructed-response exercises. At every stage, ETS and NCS closely monitor scorer selection, training, and quality.

Using the Image-Based System

The image scoring system was designed to accommodate NAEP's special needs while eliminating many of the complexities in paper-based training and scoring. First used in the 1994 assessment, the image scoring system allows scorers to assess and score student responses on line. To do this, student response booklets are scanned, constructed responses are digitized, and the images are stored for presentation on a large computer monitor. The range of possible scores for an item also appears on the display, and scorers click on the appropriate button for quick and accurate scoring.

Developed by NCS, the system facilitates the training and scoring process by electronically distributing responses to the appropriate scorers and by allowing ETS and NCS staff to monitor scorer activities consistently, identifying problems as they occur and implementing solutions expeditiously.

The system enhances scoring reliability by providing tools to monitor the accuracy of each scorer and allows scoring supervisors to create calibration sets that can be used to prevent drift in the scores assigned to questions. This tool is especially useful when scoring large numbers of responses to a question, as occurs in state NAEP, which often has more than 30,000 responses per question. The ability to prevent drift and monitor potential problems while scorers evaluate the same question for a long period is crucial to maintaining the high quality of scoring.

The image scoring system allows all responses to a particular exercise to be scored continuously until the item is finished. In an assessment such as NAEP, which utilizes a balanced incomplete block (BIB) design (see question 20 for more detail), grouping all student responses to a single question and working through the entire set of responses improves the validity and reliability of scorer judgments.

Ensuring Rater Reliability

Rater reliability refers to the consistency with which individual scorers assign a score to a question. This consistency is critical to the success of NAEP, and ETS and NCS employ three methods for monitoring reliability.

In the first method, called *backreading*, scoring supervisors review each scorer's work to confirm that the scorer applies the scoring criteria consistently across a large number of responses and that the individual does so consistently across time. Scoring supervisors evaluate

approximately 10 percent of each scorer's work in this process.

In the second method, each group of scorers performs daily *calibration* scoring so scoring supervisors can make sure that drift does not occur. Any time scorers have taken a break of more than 15 minutes (e.g., after lunch, at the start of the workday), they score a set of calibration papers that reinforces the scoring criteria.

Last, *interrater reliability* statistics confirm the degree of consistency and reliability of overall scoring, which is measured by scoring a defined percentage of the responses a second time and comparing the first and second scores.

Consistent performance among scorers is paramount for the assessment to produce meaningful results. Therefore, ETS and NCS have designed the image scoring system to allow for easy monitoring of the scoring process, early identification of problems, and flexibility in training and retraining scorers.

Measuring trends in student achievement, whether short or long term, involves special scoring concerns. To maintain a trend, scorers must train using the same materials and procedures from previous assessment years. Furthermore, reliability rates must be monitored within the current assessment year, as well as across years.

Despite consistent scoring standards and extensive training, experience shows that some discrepancies in scoring may occur between different assessment years. Thus, a random sample of 20 to 25 percent of the responses from the prior assessment is systematically interspersed among current responses for rescoring. The results are used to determine the degree of scoring agreement between the current and previous assessments, and, if necessary, current assessment results are adjusted to account for any differences.

Documenting the Process

All aspects of scoring students' constructed responses are fully documented. In addition to warehousing the actual student booklets, NCS keeps files of all training materials and reliability reports. NCS records in its scoring reports all the procedures used to assemble training packets, train scorers, and conduct scoring. These scoring reports also include all methods used to ensure reader consistency, all reliability data, and all quality control measures. ETS also summarizes the basic scoring procedures and outcomes in its technical report.

Question: 11

How does NAEP analyze the assessment results?

Answer

Before the data are analyzed, responses from the groups of students assessed are assigned sampling weights to ensure that their representation in NAEP results matches their actual percentage of the school population in the grades assessed.

Based on these sampling weights, the analyses of national and state NAEP data are conducted in two major phases for most subjects—scaling and estimation. During the scaling phase, item response theory (IRT) procedures are used to estimate the measurement characteristics of each assessment question. During the estimation phase, the results of the scaling are used to produce estimates of student achievement. Subsequent analyses relate these achievement results to the background variables collected by NAEP. Because IRT scaling is inappropriate for some groups of NAEP items, results are sometimes reported separately for each task or for each group of highly related tasks in the assessment.

NAEP data are extremely important in terms of the cost to obtain them and the reliance placed on the reports that use them. Therefore, the scaling and analysis of these data are carefully conducted and include extensive quality control checks.

FURTHER DETAILS

Weighting

Responses from the groups of students are assigned sampling weights to adjust for oversampling or undersampling from a particular group. For instance, census data on the percentage of Hispanic students in the entire student population are used to assign a weight that adjusts the NAEP sample so it is representative of the nation. The weight assigned to a student's responses is the inverse of the probability that the student would be selected for the sample.

When responses are weighted, none are discarded, and each contributes to the results for the total number of students represented by the individual student assessed. Weighting also adjusts for various situations such as school and student nonresponse because data cannot be assumed to be randomly missing. All NAEP analyses described below are conducted using these sampling weights.

Scaling and Estimation

NAEP uses IRT methods to produce score scales that summarize the results for each content area. Group-level statistics such as average scores or the percentages of students exceeding specific score points are the principal types of results reported by NAEP. However, NAEP also reports the results of various analyses, many of which examine the relationship among these group-level statistics and important demographic, experimental, and instructional variables.

Because of the reporting requirements for the national and state assessments and because of the large number of background variables associated with each assessment, thousands of analyses must be conducted. The procedures NAEP uses for the analyses were developed to produce accurate results while limiting the testing burden on students. Furthermore, these procedures provide data that can be readily used in secondary analyses.

The following procedures are used to generate scale-score data files suitable for analysis:

- After the computer files containing the student responses have been received, all cognitive and noncognitive questions are subjected to an extensive quality control analysis. An item analysis is then conducted on all of the questions. Project staff review the item analysis results, searching for anomalies that may signal unusual results or errors in creating the database. Simultaneously, each cognitive question is examined for differential item functioning (DIF). DIF analyses identify questions on which the scores of different subgroups of students who are at the same ability level differ significantly. Items showing DIF are examined by experts for potential bias.
- After the item and DIF analyses have been completed, the IRT scaling phase begins. IRT scaling provides estimates of item parameters (e.g., difficulty, discrimination) that define the relationship between the item and the underlying variable measured by the test. Parameters of the IRT model are estimated for each question, with separate scales being established for each predefined content area specified in the assessment framework. For example, in 2000 the reading assessment for grade 4 will have two scales describing reading purposes. Mathematics will have five content strands, and science will have scales for the three fields of science. Because the item parameters determine how each question is represented in the content-area scales, the psychometric staff carefully verify that the IRT scaling model provides an acceptable representation of the responses to the questions. In particular, they examine the fit of the model, by question, for the national and state assessments. Item-parameter estimations for state and national data are performed separately because the data collection processes for these assessments differ.
- During the estimation phase, plausible values (see below) are used to characterize content-area scale scores for students participating in the assessment. To keep student burden to a minimum, NAEP administers few assessment items to each student—too few to produce accurate content-related scale scores for each student. To account for this, for each student NAEP generates five possible content-related scale scores that represent selections from the distribution of content-related scale scores of students with similar backgrounds who answered the assessment items the same way. The plausible-values technology is one way to ensure that the estimates of the average performance of student populations and the estimates of variability in those estimates are more accurate than those determined

through traditional procedures, which estimate a single score for each student. During the construction of plausible values, careful quality control steps ensure that the subpopulation estimates based on these plausible values are accurate. Plausible values are constructed separately for each national sample and for each jurisdiction participating in the state assessment.

- As a final step in the analysis process, the results of assessments involving a year-to-year trend or a state component are linked to the scales for the related assessments. For national NAEP, results are linked to the scales used in previous assessments of the same subject. For state NAEP, results for the current year are linked to the scales for the nation. Linking scales in this way enables state and national trends to be studied. Comparing the scale distributions for the scales being linked determines the adequacy of the linking function, which is assumed to be linear.

Plausible Values

NAEP's assessment frameworks call for comprehensive coverage of each of the various subject areas—mathematics, science, reading, writing, civics, the arts, and others. In theory, given a sufficient number of items in a content area (a single scale within a subject-matter area), performance distributions for any population could be determined for that content area. However, NAEP must minimize its burden on students and schools by keeping assessment time brief. To do so, NAEP breaks up any particular assessment into a dozen or more blocks, consisting of multiple items, and administers only two or three blocks of items to any particular student.

This limitation results in any given student responding to only a small number of assessment items for each content area. As a result, the performance of any particular student cannot be measured accurately. The impact of this student-level imprecision has two important consequences: First, NAEP cannot report the proficiency of any particular student in any given content area (see Question 4); and second, traditional statistical methods that rely on point estimates of student proficiency become inaccurate and ineffective.

Unlike traditional standardized testing programs, NAEP must often change its test length, test difficulty, and balance of content to provide policymakers with current, relevant information. To accommodate this flexibility, NAEP uses methods that permit substantial updates between assessments but that remain sensitive enough to measure small, real changes in student performance. The use of IRT provides the technique needed to keep the underlying content-area scales the same, while allowing for variations in test properties such as changes in test length, minor differences in item content, and variations in item difficulty. NAEP estimates IRT parameters using the technique of marginal maximum likelihood, a statistical methodology. Estimations of NAEP scale score distributions are based on an estimated distribution of possible scale scores, rather than point estimates of a single scale score. This approach allows NAEP to produce accurate and statistically unbiased estimates of population characteristics that properly account for the imprecision in student-level measurement.

Marginal maximum likelihood methods are not well known or easily available to secondary analysts of NAEP data. Since most standard statistical packages provide only statistical methods that rely on point estimates of student proficiencies, rather than estimates of distributions, as the basis of their calculations, secondary analysts need an analog of point estimates that can function well with standard statistical software. For this reason, NAEP uses the plausible-values methodology as a workable alternative for secondary analysts.

Essentially, plausible-values methodology represents what the true performance of an individual might have been, had it been observed, using a small number of random draws from an empirically derived distribution of score values based on the student's observed responses to assessment items and on background variables. Each random draw from the distribution is considered a representative value from the distribution of potential scale scores for all students in the sample who have similar characteristics and identical patterns of item responses. The draws from the distribution are different from one another to quantify the degree of precision (the width of the spread) in the underlying distribution of possible scale scores that could have caused the observed performances.

The NAEP plausible values function like point estimates of scale scores for many purposes, but they are unlike true point estimates in several respects. First, they differ from one another for any particular student, and the amount of difference quantifies the spread in the underlying distribution of possible scale scores for that student. Secondary analysts must analyze the spread among the plausible values and must not analyze only one of them as if it were a true student scale score. Second, the plausible-values methodology can recover any of the potential interrelationships among score scales and subpopulations defined by background variables that have been built into the plausible values when they were generated. Although NAEP builds a great many background variables into the plausible value estimation, the relationships of any new variables (those not incorporated into the generation of the plausible values) to student scale scores may not be accurately estimated. Because of the plausible-values approach, secondary researchers can use the NAEP data to carry out a wide range of analyses.

Summary

The NAEP scaling and estimation procedures yield unbiased estimates whose quality is ensured through numerous quality control steps. NAEP uses IRT so that NAEP staff and secondary analysts can efficiently complete extensive, detailed analyses of the data. Plausible-values scaling technology enables NAEP to conduct second-phase analyses and report these results in various publications such as the *NAEP 1998 Reading Report Card for the Nation and the States*.

Question: 12

How does NAEP ensure the comparability of results among the state assessments and between the state and national assessments?

Answer

NAEP data are collected using a closely monitored and standardized process. The tight controls that guide the data collection process help ensure the comparability of the results generated for the main and the state assessments.

Main and state NAEP use the same assessment booklets, and they are administered during overlapping times. Although the administration processes for the assessments differ somewhat, statistical equating procedures that link the results from main and state NAEP to a common scale further ensure comparability. Comparing the distributions of student ability in both samples confirms the accuracy of this process and justifies reporting the results from the national and state components on the same scale.

FURTHER DETAILS

Equating Main and State Assessments

State NAEP enables each participating jurisdiction to compare its results with those for the nation and with those for the region of the country where it is located. However, before these comparisons can be made, data from the state and main assessments must be scaled separately for the following reasons:

- The assessments use different administration procedures (Westat staff collect data for main NAEP, whereas individual jurisdictions collect data for state NAEP).
- Motivational differences may exist between the samples of students participating in the main and state assessments.

For meaningful comparisons, results from the main and state assessments must be equated so they can be reported on a common scale. Equating the results depends on those parts of the main and state samples that represent a common population. Because different individuals participate in the national and state assessments of the same subject, two independent samples from the entire population are drawn from each grade assessed. These samples consist of the following:

- students tested in the national assessment who come from the jurisdictions participating in the state NAEP (called the state comparison sample, or SCS); and
- the aggregation of all students tested in the state NAEP (called the state aggregate sample, or SAS).

For the NAEP 2000 science and mathematics assessments, equating and scaling

national results with state results will occur through the common populations of fourth- and eighth-grade public school students. Separate scales will be developed for the two components and subsequently linked by equating the scale-score means and standard deviations for the SCS and SAS samples.

Verifying Comparability

After the scales have been linked, additional analyses verify the comparability of the scales for the state and national results. The verification process includes comparing the shapes of the distributions for the SCS and SAS samples. For this comparison, each scale is divided into even intervals, and the percentages of students whose scores fall into each interval are estimated. If these distributions are similar in shape, the percentages in any given interval will be similar on both scales, and the linking will have produced comparable scales. For the assessments that have occurred since this procedure was introduced, the distributions for the two components have proved very similar.

Other checks are conducted to verify the comparability. For a more detailed explanation of the linking process, see the *Technical Report of the 1996 State Assessment Program in Science*.

Related Questions:

Question 19: *How many schools and students participate in NAEP? When are the data collected during the school year?*

Question 21: *What are NAEP's procedures for collecting data?*

Question: 13

What types of reports does NAEP produce? What reports are planned for the 1999 and 2000 assessments?

Answer

NAEP has developed an information system that provides various national and local audiences with the results needed to help them monitor and improve the educational system. To have maximum utility, NAEP reports must be clear and concise, and they must be delivered in a timely fashion.

NAEP has produced a comprehensive set of reports for the 1998 assessments in reading, writing, and civics, which are targeted to specific audiences. The audiences interested in NAEP results include parents, teachers, school administrators, legislators, and researchers. Targeting each report to a segment of these interested audiences increases its impact and appeal. Selected NAEP reports are available electronically on the World Wide Web (http://nces.ed.gov/nationsreportcard), which makes them more accessible. The 2000 reports in mathematics and science and grade 4 reading will resemble those for the 1998 assessments.

FURTHER DETAILS

Reports for Different Audiences

NAEP reports are technically sound and address the needs of various audiences. For the 2000 assessments, NAEP plans to produce the following reports, most of which will be placed on the National Center for Education Statistics (NCES) Web site (http://nces.ed.gov/nationsreportcard).

NAEP Report Cards address the needs of national and state policymakers and present the results for selected demographic subgroups defined by variables such as gender, race or ethnicity, and parents' highest level of education.

Highlights Reports are nontechnical reports that directly answer questions frequently asked by parents, local school board members, and members of the concerned public.

Instructional Reports, which include many of the educational and instructional materials available from NAEP assessments, are designed for educators, school administrators, and subject-matter experts.

State Reports, one for each participating state, are intended for state policymakers, state departments of education, and chief state school officers. Customized reports will be produced for each jurisdiction that participates in the NAEP 2000 state mathematics and science assessments, highlighting the results for that jurisdiction. Mathematics results will be reported at the state level for the third time since 1992, and science results will be reported at the state level for the second time. The NAEP *2000 State Reports* will build on the computer-generated

reporting system that has been used successfully since 1990. As with past state assessments, state testing directors and state NAEP coordinators will help produce the NAEP *2000 State Reports*.

Cross-State Data Compendia, first produced for the state reading assessment in 1994, are designed for researchers and state testing directors. They serve as reference documents that accompany other reports. The *Compendia* present state-by-state results for the variables discussed in the *State Reports*.

Trend Reports describe patterns and changes in student achievement as measured through the long-term trend assessments in mathematics, science, reading, and writing. These reports present trends for the nation and for selected demographic subgroups defined by variables such as race or ethnicity, gender, region, parents' highest level of education, and type of school.

Focused Reports explore in-depth questions with broad educational implications. They provide information to educators, policymakers, psychometricians, and interested citizens.

Summary Data Tables present extensive tabular summaries based on background data from the student, teacher, and school questionnaires. A new Web tool for the presentation of these data was introduced in conjunction with the data from the 1998 assessment. The NAEP Summary Data Tables Tool is designed to permit easy access to NAEP results. The tool enables users to customize tables to more easily examine desired results. Users can also print tables and extract them to spreadsheet and word processing programs. The tool is available from the NAEP Web site (http://nces.ed.gov/nationsreportcard) and will also be available on CD–ROM.

Technical Reports document all details of a national or state assessment, including the sample design, instrument development, data collection process, and analysis procedures. *Technical Reports* only provide information about how the results of the assessment were derived; they do not present the actual results. One technical report will describe the entire 1998 NAEP, including the national assessments, the state reading assessment, and the state writing assessment. *Technical Reports* are also planned for the 1999 assessment and the 2000 assessment.

In addition to producing these reports, NAEP provides states and local school districts with continued service to help them better understand and utilize the results from the assessments. The process of disseminating and using NAEP results is continually examined to improve the usefulness of these reports.

Question: 14

What contextual background data does NAEP provide?

Answer

In addition to testing cognitive abilities, NAEP collects information from participating students, teachers, and principals about hundreds of contextual background variables regarding student, teacher, and school characteristics; instructional practices; and curricula. When developing the questionnaires used to gather this information, NAEP ensures that the questions do not infringe on respondents' privacy, that they are grounded in research, and that the answers can provide information relevant to the debate about educational reform.

The questionnaires appear in separately timed blocks of questions in the assessment booklets, such as the student questionnaires, or, like the teacher, school, and SD/LEP (students with disabilities or limited English proficiency) questionnaires, they are printed separately. Four general sources provide context for NAEP results:

- student questions, which examine background characteristics and subject area experience;
- teacher questionnaires, which gather data on background and training and classroom-by-classroom information;
- school questionnaires; and
- SD/LEP questionnaires.

These questionnaires were developed using a framework and process similar to that used for the cognitive questions. This process included reviews by external advisory groups, field testing, and additional reviews.

For the main and state NAEP, the student questions appear in noncognitive blocks. The background characteristic questions vary somewhat by grade level within a subject, and the subject area experience questions differ slightly by grade level within a subject. Unlike the cognitive blocks, these noncognitive blocks do not differ among the assessment booklets for a given grade and subject.

The teacher questionnaires vary based on subject area and may differ by grade level. The school questionnaires are completed by a school official for each grade of students participating in the assessment.

FURTHER DETAILS

Student Questionnaires

Student answers to background questions are used to gather information about factors such as race or ethnicity, school attendance, and academic expectations. Answers on those questionnaires also provide information about factors believed to influence academic performance, including homework habits, the language spoken in the home, and the quantity of reading materials in the home. Because many of these questions document changes that occur over time, they remain unchanged over assessment years.

Student subject area questions gather three categories of information: time spent studying the subject, instructional experiences in the subject, and attitudes toward and perceptions about the subject and the test. Because these questions are specific to each subject area, they can probe in some detail the use of specialized resources such as calculators in mathematics classes.

Teacher Questionnaires

To provide supplemental information about the instructional experiences reported by students, the teacher for the subject in which students are being assessed completes a questionnaire about instructional practices, teaching background, and related information.

Part I of the teacher questionnaire, which covers background and general training, includes questions concerning race or ethnicity, years of teaching experience, certifications, degrees, major and minor fields of study, course work in education, course work in specific subject areas, the amount of in-service training, the extent of control over instructional issues, and the availability of resources for the classroom.

Part II of the teacher questionnaire, which covers training in the subject area and classroom instructional information, contains questions concerning the teacher's exposure to issues related to the subject and the teaching of the subject. It also asks about pre- and in-service training, the ability level of the students in the class, the length of homework assignments, use of particular resources, and how students are assigned to particular classes.

School Questionnaires

The school questionnaire is completed by the principal or another official of the school. This questionnaire asks about the background and characteristics of the school, including the length of the school day and year, school enrollment, absenteeism, dropout rates, size and composition of the teaching staff, tracking policies, curricula, testing practices, special priorities, and schoolwide programs and problems. This questionnaire also collects information about the availability of resources, policies for parental involvement, special services, and community services.

SD/LEP Questionnaire

The SD/LEP questionnaire is completed by teachers of those students who were selected to participate in NAEP and who were classified as SD or LEP, or who had Individual Education Plans (IEPs) or equivalent classification. The SD/LEP

questionnaire gathers information about the background and characteristics of each student and the reason for the SD/LEP classification. For a student classified as SD, the questionnaire requests information about the student's functional grade level, mainstreaming, and special education programs. For a student classified as LEP, questions ask about the student's native language, time spent in special education and language programs, and the level of English language proficiency.

NAEP policy states that if any doubt exists about a student's ability to participate, the student should be *included* in the assessment. Beginning with the 1996 assessments, NAEP has allowed more accommodations for both categories of students.

Related Question:

Question 7: *How does NAEP accommodate students with disabilities and students with limited English proficiency?*

Question: 15

How can educators use NAEP resources such as frameworks, released questions, and reports in their work?

Answer

NAEP materials such as frameworks, released questions, and reports have many uses in the educational community. For instance, states have considered NAEP frameworks when revising their curricula. Also, released constructed-response questions and their corresponding scoring guides have served as models of innovative assessment practices.

NAEP findings are reported in many publications specifically targeted to educators. Furthermore, NAEP staff host seminars to discuss NAEP results and their implications.

FURTHER DETAILS

NAEP Frameworks

NAEP frameworks present and explain what experts in a particular subject area consider important. Each framework outlines the subject, often providing examples, in ways that may give teachers and curriculum planners new perspectives about their fields.

Frameworks frequently provide theoretical information about problem solving through their descriptive classifications of cognitive levels (e.g., in the mathematics framework, conceptual understanding, procedural knowledge, and problem solving). Educators can relate these cognitive levels to various subject content areas and evaluate how classroom instruction and assessment focus on each cognitive level. For example, an instructor may study the mathematics framework and see that most of his or her instruction addresses only procedural knowledge. The instructor can then include more problems at a higher cognitive level, perhaps following examples suggested in the framework.

Several states have used the framework publications to review their NAEP results and develop recommendations for teachers. Furthermore, NAEP staff members who are familiar with particular frameworks have lent their expertise to state curriculum committees.

Released NAEP Questions

After each assessment, the National Center for Education Statistics (NCES) releases nearly one-third of the questions, making copies of them available to the interested public. The packages containing the released questions include answer keys, content and process descriptions, and information about the percentages of students who answered the questions correctly.

Released questions are also available on the NAEP Web site (http://nces.ed.gov/nationsreportcard). The sample questions tool displays test questions, along with sample student responses and

scoring guides from the assessment. The test questions can be downloaded and printed directly from the Web site.

Released questions often serve as models for teachers who wish to develop their own classroom assessments. One school district used released NAEP reading questions to design a districtwide test, and another school district used scoring guides for released reading questions to train its teachers in how to construct scoring guides.

NAEP Reports

NAEP reports such as the focus report on mathematical problem solving provide teachers with useful information. NAEP staff have also conducted seminars for school districts across the country to discuss NAEP results and their implications at the local level. In 1996, NCES began placing NAEP reports and almanacs on its World Wide Web site (http://nces.ed.gov/nationsreportcard) for viewing, printing, and downloading. Web access should increase the utility of NAEP results.

Related Question:

Question 4: *Can parents examine the questions NAEP uses to assess student achievement? Can parents find out how well their children performed in the NAEP assessment? Why are NAEP questions kept confidential?*

Question: 16

How are NAEP data and assessment results used to further explore education and policy issues? What technical assistance does NAEP provide?

Answer

The National Center for Education Statistics (NCES) grants members of the educational research community permission to use NAEP data. Educational Testing Service (ETS) provides technical assistance, either as a public service or under contract, in using these data.

NAEP results are provided in formats that the general public can easily access. Tailored to specific audiences, NAEP reports are widely disseminated. Since the 1994 assessment, reports and almanacs have been placed on the World Wide Web to provide even easier access.

FURTHER DETAILS

NAEP Data

Because of its large scale, the regularity of its administration, and its rigid quality control process for data collection and analysis, NAEP provides numerous opportunities for secondary data analysis. NAEP data are used by researchers who have many interests, including educators who have policy questions and cognitive scientists who study the development of abilities across the three grades assessed by NAEP.

World Wide Web Presence

Beginning with the 1994 assessment, NCES began placing NAEP reports and almanacs on its World Wide Web site (http://nces.ed.gov/nationsreportcard) for viewing, printing, and downloading.

Software and Data Products

NAEP has developed products that support the complete dissemination of NAEP results and data to many analysis audiences. ETS began developing these data products for the 1990 NAEP, adding new capabilities and refinements in subsequent assessments.

In addition to the user guides and a version of the NAEP database for secondary users on CD-ROM, these other products are available:

- the NAEP Summary Data Tables Tool for searching, displaying, and customizing cross-tabulated variable tables (available on the NAEP Web site and on CD-ROM); and
- the NAEP Data Tool Kit, including *NAEPEX*, a data extraction program for choosing variables, extracting data, and generating SAS and SPSS control statements, and analysis modules for cross-tabulation and regression that work with SPSS and Excel (available on disk).

ETS and NCES conduct workshops on how to use these products to promote secondary analyses of NAEP data.

NAEP Technical Assistance

Seminars

Frequently, NCES offers a four-day seminar—the NCES Advanced Studies Seminar on the Use of the NAEP Database for Research and Policy Discussion. This seminar stimulates interest in using NAEP data to address educational research questions, enhances participants' understanding of the methodological and technological issues relevant to NAEP, and demonstrates the steps necessary for conducting accurate statistical analyses of NAEP data. In addition to offering formal and hands-on instruction, the seminar helps participants learn about and work with currently available software packages specifically designed for NAEP analyses.

Special Analyses

Under the cooperative agreement with NCES, ETS develops software for secondary users of NAEP data and regularly provides them with technical assistance either by supplying information about data characteristics or by contracting with them to run analyses.

Special Studies

NAEP meets requests for the empirical investigation of educational issues. Nearly every NAEP assessment contains a special study that gathers specific information expressly requested by the public and NCES.

State-Level Forum for Discussing Educational Issues and Policy

The NAEP NETWORK provides a state-level forum for educational concerns. Participants in these meetings include testing directors; NAEP coordinators from individual states, territories, and other jurisdictions; and representatives from nonpublic school organizations and associations. The NAEP NETWORK also offers information about upcoming assessments and enables those involved in state NAEP to offer their input.

Related Question:

Question 13: *What types of reports does NAEP produce? What reports are planned for the 1999 and 2000 assessments?*

Question: 17

Can NAEP results be linked to other assessment data?

Answer

In recent years there has been considerable interest among education policymakers and researchers in linking NAEP results to other assessment data. Much of this interest has been centered on linking NAEP to international assessments. The 1992 NAEP mathematics assessment results were successfully linked to those from the International Assessment of Educational Progress (IAEP) of 1991, and the 1996 grade 8 NAEP assessments in mathematics and science have been linked to the results of the Third International Mathematics and Science Study (TIMSS) of 1995. Also, a number of activities have focused on linking NAEP to state assessment results. Promoting linking studies with international assessments and assisting states and school districts in linking their assessments to NAEP are key aspects of the National Assessment Governing Board's (NAGB's) policy for redesigning NAEP.

FURTHER DETAILS

Linking NAEP to International Assessments

The International Assessment of Educational Progress (IAEP)

Pashley and Phillips (1993) investigated linking mathematics performance on the 1991 IAEP to performance on the 1992 NAEP. In 1992, they collected sample data from U.S. students who were administered both instruments. (Colorado drew a large enough sample to compare itself with all 20 countries that participated in IAEP.)

The relation between mathematics proficiency in the two assessments was modeled using regression analysis. This model was then used for projecting IAEP scores from non-U.S. countries onto the NAEP scale.

The authors of the study considered their results very encouraging. The relation between the IAEP and NAEP assessments was relatively strong and could be modeled well. However, as the authors pointed out, the results should be considered only in the context of the similar construction and scoring of the two assessments. Thus, they advised that other studies should be initiated cautiously, even though the path to linking assessments was better understood.

The Third International Mathematics and Science Study (TIMSS)

In 1989, the United States expressed an interest in international comparisons, especially in mathematics and science. That year, the National Education Summit adopted goals for education. Goal 4 states that American students shall be first in the world in mathematics and science achievement by the year 2000. Since that pronouncement, various approaches have been suggested for collecting the data that could help monitor progress toward that goal.

The 1995 TIMSS presented one of the best opportunities for comparison. The data from this study became available at approximately the same time as the NAEP data for the 1996 mathematics and science assessments. Because the two assessments were conducted in different years and no students responded to both assessments, the regression procedure that linked the NAEP and IAEP assessments could not be used. Therefore, the results from the NAEP and TIMSS assessments were linked by matching their score distributions (Johnson & Owen, 1998). A comparison of linked grade 8 results with actual grade 8 results from states that participated in both assessments suggested that the link was working acceptably.

A research report based on this linking (Johnson, Siegendorf, & Phillips, 1998) provides comparisons of the mathematics and science achievement of each U.S. jurisdiction that participated in the state NAEP with that of each country that participated in TIMSS. However, as was the case with the IAEP link, these comparisons need to be interpreted cautiously.

The same linking approach did not produce satisfactory results at grade 4, and no comparisons at this grade have been reported. Studies to date have yielded no information as to why the distribution matching method produced acceptable results at one grade but unacceptable results at the other. The National Center for Education Statistics (NCES) plans to repeat the linking of NAEP and TIMSS as part of the NAEP 2000 assessment. However, in this linking effort, a sample of students will be administered both the NAEP and TIMSS assessments. As a result, regression-based procedures like those used in the NAEP-to-IAEP linking can be employed. It is hoped that the use of these procedures will provide useful linkages at all grades.

Linking NAEP to State Assessments

One way in which NAEP can be made most useful to state education agencies is by providing a benchmark against which they can compare the results of the census assessments they carry out in their schools. If the state assessment scores are mapped onto the NAEP scale, then states will be in a position to make stronger statements about the implications of performance on their state assessment than would otherwise be possible. If the projections are valid, schools and districts can compare their scores not only to the state as a whole but also to the entire nation and, based on linkage of NAEP to international assessments, to other major countries in the world.

Because this capability is valuable, states have examined alternative methods for developing linkages between their assessments and NAEP assessments. Many methods may not produce valid results because they do not link the operational state assessment to the operational NAEP on which the scales are defined. Varying the context in which an item or a test is administered varies the performance of students on that item, so that a valid linkage must be based on linking scores on actual administrations of NAEP and state assessments.

Building on the earlier work of Linn (1993), Bloxom et al. (1995), and Williams et al. (1995), McLaughlin (1998a) explored the feasibility and validity of regression-based linkings based on matching state assessment scores of students to NAEP performance records. Using the 1996 state NAEP grade 4 and 8 mathematics

assessments in four states, he found that: (a) it is feasible to develop the linkage of student records without violating either NAEP or state assessment confidentiality assurances, and (b) in three of the four states, acceptably accurate projections of group-level NAEP scores and percentages at achievement levels could be obtained.

McLaughlin (1998b) found that in order to be neutral (i.e., so that comparisons based on projected NAEP scores lead to the same conclusions as comparisons based on actual NAEP scores), it was necessary that the regression models include explicit terms for school mean scores, as well as individual student scores, and explicit terms for demographic measures. Like others (Linn & Kiplinger, 1993; Shepard, 1997), he also found that projection functions did not necessarily generalize across years.

As part of the exploration of linkage, McLaughlin developed detailed guidelines for validating score projections and for conveying the appropriate level of uncertainty in statements based on projections. McLaughlin's methodology for linking and for the development and validation of projections is being made available to states, including both the development and evaluation of the linkage database and the execution of a specified series of analyses to derive the appropriate lining function and evaluate its precision, neutrality, and stability.

Question: 18

Who are the students assessed by NAEP?

Answer

NAEP does not, and is not designed to, report on the performance of individual students. Rather, it assesses specific populations of in-school students or subgroups of these populations, reporting on their performance in selected academic areas. NAEP results are based on samples of these student populations of interest.

NAEP assesses representative samples of students in certain grades or at certain ages in public and nonpublic schools in the United States. For the main NAEP assessment, students are selected from grades 4, 8, and 12 in public and nonpublic schools. For the state assessments, the number of subjects and grades selected varies, depending on available funding. (State-level samples have included public and nonpublic school students since 1994. In 2000, however, nonpublic schools will not be included in the state-level samples.) For the long-term trend assessments, students at grades 4, 8, and 11 are sampled for writing, and students at ages 9, 13, and 17 are sampled for science, mathematics, and reading. Since 1984, NAEP has been administered every two years. Beginning in 2000, NAEP will be administered annually.

FURTHER DETAILS

Sampling Students

NAEP does not report on the performance of individual students. Instead, the main and state assessments are administered every two years to representative samples of students at certain ages or in certain grades in public and nonpublic schools. Assessment results are based on the performance of students in these samples. For this reason, NAEP has developed complex sampling designs that produce precise estimates of student performance and maximize the information available given scarce resources such as students' and teachers' time.

The number of children in any particular grade in the United States is approximately 3.5 million, and the number of students selected for NAEP's national samples for any particular grade and subject is 7,000 or more. Therefore, NAEP's sample consists of approximately 0.2 percent of the student population for each grade and subject.

Although only a very small percentage of the student population in each grade is assessed, NAEP estimates are accurate because they depend on the absolute number of students participating, not on the relative proportion of students. Thus, all or nearly all of the schools and students selected must participate in the assessment to ensure that the NAEP sample is truly representative of the nation's student population.

Ensuring Representative Samples

As the Nation's Report Card, NAEP must report accurate results for populations of students and subgroups of these populations (e.g., minority students or students attending nonpublic schools). To ensure accurate results, the relatively small samples of students selected for the NAEP assessments must be truly representative of the entire student population.

Every school has some known chance of being selected for the sample. Within a selected school, all students within a participating grade have an equal chance of being selected. The probability of students and schools being selected into the sample varies based on factors such as grade, subject, public and nonpublic status, and so on.

Uncertainty about the validity of the results can arise, however, if selected schools decide not to participate, if selected students are absent or refuse to participate, or if schools or students volunteer to join the NAEP sample. For this reason, NAEP encourages the participation of all those selected, but is not able to accept volunteers.

Stratification

A multistage design that relies on stratification (i.e., classification into groups having similar characteristics) is used to choose samples of student populations. To ensure an accurate representation, the samples are randomly selected from groups of schools that have been *stratified* by variables such as region of the country, extent of urbanization, percentage of minority enrollment, and median household income.

The samples drawn for state assessments are separate from those drawn for national assessments. However, because state NAEP assessments are voluntary, some states may choose not to participate. Therefore, the aggregate of the state samples may not be a representative sample of the nation.

For main and long-term trend NAEP assessments, the sampling design has three steps:

- selection of primary sampling units (PSUs), the geographic areas defined as counties or groups of counties;
- selection of schools (public and nonpublic) within the selected areas; and
- selection of students within those schools.

Stratification begins when PSUs are identified based on region of the country (Northeast, Southeast, Central, and West) and when PSUs within each region are designated as urban or rural. In some regions, PSUs are classified by the percentage of Black or Hispanic students enrolled in their school populations. Further strata may be defined using other variables such as median household income, education level of residents over 25 years of age, or other demographic characteristics. Once the selection of the PSUs is completed, the schools within each PSU are assigned a probability of selection that is proportional to the number of students per grade in each school.

For state NAEP 2000 assessments, the sampling design has two steps:

- selection of public schools within the selected areas; and
- selection of students within those schools.

To ensure that the state samples provide an accurate representation, public schools are stratified by urbanization, minority enrollment, and the results of state achievement tests or median household income.

Oversampling

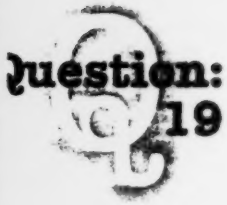
To ensure that the results reported for major subgroups of populations are accurate, *oversampling* (i.e., sampling particular types of schools at a higher rate than they appear in the population) is necessary. For example, main NAEP oversamples nonpublic schools and schools with large minority populations—which ensures that the sample contains adequate numbers of students attending nonpublic schools and students from particular racial or ethnic subgroups to accurately estimate the performance of those subgroups.

Weighting

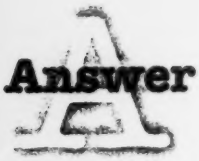
If these samples are to be representative of the population as a whole, however, the data from the students in the oversampled schools must be properly *weighted* during analysis. Weighting accounts for the disproportionate representation of certain subgroups that occurs because of oversampling. Similarly, weighting also accounts for low sampling rates that can occur for very small schools. Thus, when properly weighted, NAEP data provide results that reflect the representative performances of the entire nation and of the subpopulations of interest.

Related Question:

Question 1: *What is NAEP?*



How many schools and students participate in NAEP? When are the data collected during the school year?



National assessment samples typically include nearly 100,000 students from 2,000 schools. The state assessment samples usually include approximately 100 schools and 2,500 students per subject per grade from each participating state.

Data for the main and state NAEP are collected at overlapping times during winter. Data for the national long-term trend assessments are collected during fall for age 13 or grade 8, during winter for age 9 or grade 4, and during spring for age 17 or grade 11.

FURTHER DETAILS

Sample Selection

Each assessment year, NAEP selects national samples that represent the entire population of U.S. students in grades 4, 8, and 12 for the main assessment and the entire population of students at ages 9, 13, and 17 for the long-term trend assessment. The selection process differs slightly based on whether the sample of students is needed for the main assessment, the long-term trend assessment, or the state-level assessment.

Main and State NAEP

During any given assessment year, subject areas may not be assessed at all grades. For instance, main NAEP in 2000 will assess students in mathematics and science at all three grades, but only at grade 4 in reading. State NAEP in 2000 will assess students in mathematics and science at grades 4 and 8, but it will not test twelfth-grade students in either subject.

The 1998 main assessment, which tested reading, writing, and civics at all three grade levels, required about 137,000 students, whereas the state-level assessment, which tested only two grade levels, required a total sample of about 350,000 students because of the number of states that participated. Approximately 2,000 schools took part in the 1998 main assessment, while nearly 10,000 schools contributed their students' time and other resources to the state NAEP.

Long-Term Trend NAEP

The long-term trend assessment tests the same four subjects across years, using relatively small national samples. For the long-term trend assessment, samples of students are selected by age (9, 13, and 17 years) for mathematics, science, and reading and by grade (4, 8, and 11) for writing. Based on precedent, long-term trend writing results are reported by grade, and mathematics, science, and reading long-term trend results are reported by age. In 1999, the long-term trend NAEP began to be administered every four years, but not in the same years as the main and state assessments in reading, writing, mathematics, and science.

Schools and Students Assessed

The table below shows the target sample size of schools and students needed for main and state NAEP in 2000. These components use separate, nonoverlapping samples whenever possible.

Assessment Schedules

The table on page 58 presents the typical assessment periods for the three NAEP components. The assessment schedules remain relatively constant across assessment years to permit an accurate measurement of change over time and to help ensure that the results are comparable. The long-term trend assessment is administered three times during the school year (one administration per grade), whereas the main and state NAEP assessments are administered during winter for all three grades.

State NAEP assessments are usually administered during February. Because the month-long state assessment period occurs during the middle of the two-and-one-half-month national assessment period, the effects that a time difference could produce are eliminated, making the results of the state and national assessments more comparable than if they were administered at different times during the year.

Total Schools and Students:* Target Sample Sizes for the 2000 NAEP Assessments

NAEP COMPONENTS	Total Schools	Total Students			
NATIONAL NAEP	2,500	106,500	**Grade 4**	**Grade 8**	**Grade 12**
Reading			8,000		
Mathematics			21,750	15,750	13,750
Science			15,750	15,750	15,750
STATE NAEP	average 225 per state	500,000	**Grade 4**	**Grade 8**	**Grade 12**
Mathematics			125,000	125,000	
Science			125,000	125,000	
TOTALS for all components		606,500	295,500	281,500	29,500

* Numbers are for the total samples rather than for the reporting samples; therefore, these sample sizes will be larger than the sample sizes published in reports of NAEP results.

BEST COPY AVAILABLE

Typical NAEP Assessment Schedule

Grade or Age:	FALL	WINTER	SPRING
Grade 4 (main, state) Age 9/Grade 4 (LTT*)		Main State Long-Term Trend	
Grade 8 (main, state) Age 13/Grade 8 (LTT*)	Long-Term Trend	Main State	
Grade 12 (main) Age 17/Grade 11 (LTT*)		Main	Long-Term Trend

* LTT refers to ages/grades sampled for long-term trend assessments.

Related Questions:

Question 2: *What subjects does NAEP assess? How are the subjects chosen, and how are the assessment questions determined? What subjects were assessed in 1999? What subjects will be assessed in 2000?*

Question 20: *How does NAEP use matrix sampling? What is focused BIB spiraling, and what are its advantages for NAEP?*

Question: 20

How does NAEP use matrix sampling? What is focused BIB spiraling, and what are its advantages for NAEP?

Answer

Typically, several hundred questions are needed to reliably test the many specifications of the complex frameworks that guide the NAEP assessments. Administering the entire collection of cognitive questions to each student would be far too time consuming to be practical.

Therefore, a number of test booklets are printed for a particular subject, with each booklet containing a different selection of cognitive blocks. This design, called matrix sampling, allows NAEP to assess the entire subject area within a reasonable amount of testing time. In matrix sampling, different portions from the entire pool of cognitive questions are printed in separate booklets and administered to different but equivalent samples of students. This design minimizes the assessment time required per student while allowing complete coverage of the subject being assessed.

In addition to the background questions, the NAEP test booklets contain cognitive questions that are arranged in collections of separately timed blocks. For the cognitive questions, the blocks and their sequence vary within each version of the test booklet. This planned variation of test booklets allows NAEP to generate precise estimates of student performance and maximize the information available given scarce resources such as students' and teachers' time.

The type of matrix sampling used by NAEP is called focused, balanced incomplete block (BIB) spiraling. Because of BIB spiraling, NAEP can sample enough students to obtain precise results for each question while generally consuming an average of about an hour and a half of each student's time.

The "focused" part of NAEP's matrix sampling method requires each student to answer questions from only one subject area. The "BIB" part of the method ensures that students receive different interlocking sections of the assessment forms, enabling NAEP to check for any unusual interactions that may occur between different samples of students and different sets of assessment questions. "Spiraling" refers to the method by which test booklets are assigned to pupils, which ensures that any group of students will be assessed using approximately equal numbers of the different versions of the booklet.

FURTHER DETAILS

BIB Spiraling

Matrix sampling is used to construct tests when the objectives being assessed require more time than is available to test them. Although BIB spiraling is only one form of matrix sampling, the other forms have drawbacks that make them unsuitable for NAEP.

Occasionally, NAEP may use simple matrix sampling when it matches the needs of certain assessment questions. With simple matrix sampling, separate sets of questions are confined to particular booklets. However, because each booklet contains a set of specific questions, context and order effects must be considered. In a simple matrix design, the same subject area questions would always appear in the same place in every assessment booklet. Thus, student mastery of the questions appearing at the end of the booklet may be underestimated because of fatigue or overestimated due to practice effects.

By contrast, the more sophisticated BIB method produces data that are relatively free from these placement effects. In the NAEP BIB design, the cognitive blocks are balanced. Each cognitive block appears an equal number of times in every possible position. Each cognitive block is also paired with every other cognitive block in at least one test booklet. (The NAEP BIB design varies according to subject area.)

The following table presents a simplified example of BIB spiraling based on the NAEP 1990 mathematics design. The full sample of students is divided into seven equivalent groups, and each group of students is assigned one of the seven test booklets. In this design, each cognitive block appears only once in each of three possible positions, and each block is paired once with every other block. (This example shows only the cognitive blocks, even though the test booklets also contain background questionnaire blocks.)

A Model of BIB Matrix Sampling

Booklet version	Position 1 cognitive block	Position 2 cognitive block	Position 3 cognitive block
1	A	B	D
2	B	C	E
3	C	D	F
4	D	E	G
5	E	F	A
6	F	G	B
7	G	A	C

Question 20

69

NAEP uses BIB spiraling even though this design requires a greater variety of test booklets to be printed. Furthermore, each version of the assessment booklet must appear in the sample approximately the same number of times and must be administered to equivalent subgroups within the full sample. To ensure proper distribution at assessment time, the booklets are packed in spiral order (in the above example, one each of booklets 1 through 7, then 1 through 7 again, and so on). The test coordinator randomly assigns these booklets to the students in each test administration session. Spiraled distribution of the booklets promotes comparable sample sizes for each version of the booklet, ensures that these samples are randomly equivalent, and reduces the likelihood that students will be seated within viewing distance of an identical booklet.

Question: 21

What are NAEP's procedures for collecting data?

Answer

A cooperative agreement between the National Center for Education Statistics (NCES) and Westat specifies the sampling and data collection operations for national and state NAEP.

Westat field staff, who receive extensive training, administer the national assessment. Although each participating state is responsible for data collection for the state NAEP, Westat ensures uniformity of procedures across states through detailed procedural manuals, training, supervision, and quality control monitoring.

The complex process by which NAEP collects data is monitored closely. The tight control on this process contributes to the quality—and thus to the comparability—of the main and state assessments and their results.

FURTHER DETAILS

Organization and Supervision of Data Collection

NAEP relies heavily on the support of school administrators and staff, and obtaining the cooperation of the selected schools requires substantial time and energy because participation in NAEP is voluntary. A series of mailings that includes letters to the chief state school officers and district superintendents notifies the sampled schools of their selection. Additional informational materials are sent and procedures are explained at introductory meetings.

Westat is responsible for the following field administration duties:

- selecting the sample of schools and students;
- developing the administration procedures, manuals, and materials;
- hiring and training staff to conduct the assessments (for main NAEP);
- training state personnel to conduct assessments (for state NAEP); and
- conducting an extensive quality assurance program.

For the main and long-term trend assessments, Westat hires and trains approximately 85 field supervisory staff to collect the data. These field supervisors and their assistants complete all associated paperwork, reducing the burden on participating schools.

For the state assessments, NAEP legislation requires each participating jurisdiction to handle data collection activities such as obtaining the cooperation of sampled schools and assigning personnel to conduct the assessment. Westat employs and trains state supervisors to work with the state-appointed coordinators who carry out the necessary organizational

tasks. The individual schools and the designated assessment administrators are responsible for preparing lists of enrolled students for the sampled grade; distributing the teacher, school, and SD/LEP (students with disabilities or limited English proficiency) questionnaires; and administering the assessment.

In addition to training local administrators, Westat ensures quality control across states by monitoring 25 percent of the sessions in states that have previously conducted a NAEP assessment and 50 percent of the sessions in states that are participating in NAEP for the first time. Security of assessment materials and uniformity of administration are high priorities for NAEP. Quality control monitors have never reported any instances of serious breaches in procedures or major problems that could jeopardize the validity of the assessment.

After each session, Westat staff interview the assessment administrators to receive their comments and recommendations. As a final quality control step, a debriefing meeting is held with the state supervisors to receive feedback that will help improve procedures, documentation, and training for future assessments.

Management of Assessment Materials

Under the direction of Educational Testing Service (ETS) staff, National Computer Systems (NCS) produces the materials needed for the NAEP assessments. NCS prints identifying bar codes and numbers for the booklets and questionnaires, preassigns the booklets to testing sessions, and prints the booklet numbers on the administration schedule. These activities improve the accuracy of data collection and assist with the spiraled distribution process.

Preassigning numbered test booklets to assessment sessions and printing the booklet numbers on the administration schedule improve the systematic protection of student confidentiality. Student names appear only on stickers that are temporarily attached to the booklets; these stickers are removed and destroyed immediately after the session. Furthermore, the administration forms are perforated so all student and teacher names can be easily removed. The forms containing the student names are never removed from the school, even after the assessments are completed, and names are never linked to student or teacher data.

NCS handles all receipt control, data preparation and processing, scanning, and scoring activities for the NAEP assessments. Using an image-processing and scoring system specially designed for NAEP, NCS scans the multiple-choice selections, the handwritten student responses, and other data provided by students, teachers, and administrators. When this image-based scoring system was introduced during the 1994 assessment, it virtually eliminated paper handling during the scoring process. The system also permits scoring reliability to be monitored on line and recalibration methods to be introduced.

Related Questions:

Question 3: *Is participation in NAEP voluntary? Are the data confidential? Are students' names or other identifiers available?*

Question 20: *How does NAEP use matrix sampling? What is focused BIB spiraling, and what are its advantages for NAEP?*

Allen, N. L., Carlson, J. E., & Zelenak, C. A. (1999). *The NAEP 1996 technical report.* Washington, DC: National Center for Education Statistics.

Allen, N. L., Jenkins, F., & Zelenak, C. A. (1997). *Technical report of the NAEP 1996 state assessment program in mathematics.* Washington, DC: National Center for Education Statistics.

Allen, N. L., Mazzeo, J., Ip, E. H. S., Swinton, S., Isham, S. P., & Worthington, L. (1995). Data analysis and scaling for the 1994 Trial State Assessment in reading. In J. Mazzeo, N. L. Allen, & D. L. Kline, *Technical report of the NAEP 1994 trial state assessment in reading* (pp. 169–219). Washington, DC: National Center for Education Statistics.

Allen, N. L., Swinton, S. S., Isham, S. P., & Zelenak, C. A. (1998). *Technical report of the NAEP 1996 state assessment program in science.* Washington, DC: National Center for Education Statistics.

Bloxom, B. P. J., Nicewander, W. A., & Yan, D. (1995). Linking to a large-scale assessment: An empirical evaluation. *Journal of Educational and Behavioral Statistics, 20,* 1–26.

Calderone, J., King, L. M., Horkay, N. (eds). (1997). *The NAEP Guide.* Washington, DC: National Center for Education Statistics.

Campbell, J. R., Reese, C. M., O'Sullivan, C., & Dossey, J. A. (1996). *NAEP 1994 trends in academic progress.* Washington, DC: National Center for Education Statistics.

Campbell, J. R., Voelkl, K. E., & Donahue, P. L. (1997). *NAEP 1996 trends in academic progress.* Washington, DC: National Center for Education Statistics.

Campbell, J. R., Voelkl, K. E., & Donahue, P .L. (1998). *NAEP 1996 trends in academic progress: Addendum.* Washington, DC: National Center for Education Statistics.

College Board. (1999). *Mathematics framework for the 1996 and 2000 National Assessment of Educational Progress.* Washington, DC: National Assessment Governing Board.

Council of Chief State School Officers. (1999a). *Science framework for the 1996 and 2000 National Assessment of Educational Progress.* Washington, DC: National Assessment Governing Board.

Council of Chief State School Officers. (1999b). *Reading framework for the National Assessment of Educational Progress: 1992–2000.* Washington, DC: National Assessment Governing Board.

Donahue, P. L., Voelkl, K. E., Campbell, J. R., & Mazzeo, J. (1999). *NAEP 1998 Reading Report Card for the Nation and the States.* Washington, DC: National Center for Education Statistics.

Green, J. L., Burke, J., & Rust, K. L. (1995). Sample design and selection. In J. Mazzeo, N. Allen, & D. Kline, *Technical report of the NAEP 1994 trial state assessment in reading* (pp. 35–69). Washington, DC: National Center for Education Statistics.

Johnson, E. G., & Owen, E. (1998). *Linking the National Assessment of Educational Progress (NAEP) and the Third International Mathematics and Science Study (TIMSS): A technical report.* Washington, DC: National Center for Education Statistics.

Johnson, E. G., Siegendorf, A., and Phillips, G. W. (1998). *Linking the National Assessment of Educational Progress (NAEP) and the Third International Mathematics and Science Study (TIMSS): Eighth-grade results.* Washington, DC: National Center for Education Statistics.

Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education, 6,* 83–102.

Linn, R. L., & Kiplinger, V. L. (1993). *Linking statewide tests to the National Assessment of Educational Progress: Stability of results.* Boulder, CO: Center for Research on Evaluation, Standards, and Student Testing.

Linn, R. L., Koretz, D., & Baker, E. L. (1996). Assessing the validity of the National Assessment of Educational Progress: NAEP Technical Review Panel white paper. Washington, DC: U.S. Department of Education.

McLaughlin, D. H. (1998a). *Study of the linkages of 1996 NAEP and state mathematics assessments in four states.* Washington, DC: National Center for Education Statistics.

McLaughlin, D. H. (1998b). *Linking state assessments of NAEP: A study of the 1996 mathematics assessment.* Paper presented to the American Educational Research Association, San Diego, CA.

National Assessment of Educational Progress. (1982). *Writing objectives: 1984 assessment.* Washington, DC: U.S. Department of Education.

National Assessment of Educational Progress. (1984). *Reading objectives: 1983–84 assessment.* Washington, DC: U.S. Department of Education.

National Assessment of Educational Progress. (1985a). *Math objectives: 1985–86 assessment.* Washington, DC: U.S. Department of Education.

National Assessment of Educational Progress. (1985a). *Science objectives: 1985–86 assessment.* Washington, DC: U.S. Department of Education.

National Assessment of Educational Progress. (1987). *Writing objectives: 1988 assessment.* Washington, DC: U.S. Department of Education.

National Assessment of Educational Progress. (1989). *Science objectives: 1990 assessment.* Washington, DC: U.S. Department of Education.

Pashley, P. J., & Phillips, G. W. (1993). *Toward world-class standards: A research study linking international and national assessments.* Princeton, NJ: Educational Testing Service.

Pellegrino, J. W., Jones, L. R., & Mitchell, K. J. (eds). (1999). *Grading the nation's report card: Evaluating NAEP and transforming the National Assessment of Educational Progress.* Washington, DC: National Academy Press.

Shepard, L.A. (1997). *Measuring achievement: What does it mean to test for robust understanding?* William H. Angoff Memorial Lecture Series. Princeton, NJ: Educational Testing Service.

Wallace, L., & Rust, K. F. (1996). Sample design. In N. Allen, D. Kline, & C. Zelenak, *NAEP 1994 technical report* (Chapter 3). Washington, DC: National Center for Education Statistics.

Williams, V. S. L., Billeaud, K., Davis, L. A., Thissen, D., & Sanford, E. (1995). Projecting to the NAEP scale: Results from the North Carolina End-of-Grade Testing Program. *Journal of Educational Measurement.*

Bock, D. B., & Zimowski, M. F. (1998). *Feasibility studies of two-stage testing in large-scale educational assessment: Implications for NAEP.* Commissioned by the NAEP Validity Studies Panel. Palo Alto, CA: American Institutes for Research.

Bourque, M. L., Campagne, A. B., & Crissman, S. (1997). *1996 science performance standards: Achievement results for the nation and the states.* Washington, DC: National Assessment Governing Board.

Campbell, J. R., & Donahue, P. L. (1997). *Students selecting stories: The effects of choice in reading assessment.* Washington, DC: National Center for Education Statistics.

Chromy, J. R. (1998). *The effects of finite sampling on state assessment sample requirements.* Commissioned by the NAEP Validity Studies Panel. Palo Alto, CA: American Institutes for Research.

Dossey, J. A., Mullis, I., & Jones, C. O. (1993). *Can students do mathematical problem solving?* Washington, DC: U.S. Department of Education.

Hawkins, E. F., Stancavage, F., & Dossey, J. A. (1998). *School policies affecting instruction in mathematics.* Washington, DC: National Center for Education Statistics.

Hedges, L. V., & Vevea, J. L. (1997). *A study of equating in NAEP.* Commissioned by the NAEP Validity Studies Panel. Palo Alto, CA: American Institutes for Research.

Jaeger, R. M. (1998). *Reporting the results of the National Assessment of Educational Progress.* Commissioned by the NAEP Validity Studies Panel. Palo Alto, CA: American Institutes for Research.

Jakwerth, P. R., Stancavage, F. B., & Reed, E. D. (1999). *An investigation of why students do not respond to questions.* Commissioned by the NAEP Validity Studies Panel. Palo Alto, CA: American Institutes for Research.

Lindquist, M. M., Dossey, J. A., & Mullis, I. (1995). *Reaching standards: A progress report on mathematics. A policy information perspective.* Princeton, NJ: Educational Testing Service.

Mislevy, R. J. (1992). *Linking educational assessments.* Princeton, NJ: Educational Testing Service.

Mitchell, J. H., Hawkins, E. F., Jakwerth, P., Stancavage, F. B., & Dossey, J. A. (1999). *Student work and teacher practices in mathematics.* Washington, DC: U.S. Department of Education.

Mullis, I. V. S. (1997). *Optimizing state NAEP: Issues and possible improvements.* Commissioned by the NAEP Validity Studies Panel. Palo Alto, CA: American Institutes for Research.

Mullis, I., Jenkins, F. L., & Johnson, G. (1994). *Effective schools in mathematics.* Washington, DC: National Center for Education Statistics.

O'Sullivan, C. Y., Weiss, A. R., & Askew, J. M. (1998). *Students learning science: A report on policies and practices in U.S. schools.* Washington, DC: National Center for Education Statistics.

Pearson, D. P., & Garavaglia, D. R. (1997). *Improving the information value of performance items in large-scale assessments.* Commissioned by the NAEP Validity Studies Panel. Palo Alto, CA: American Institutes for Research.

almanac. A comprehensive collection of tables of NAEP results.

assessment session. The period of time during which a test booklet is administered to students.

background questionnaires. The instruments used to collect information about student demographics and educational experiences.

bias. In statistics, the difference between the expected value of an estimator and the population parameter being estimated. If the average value of the estimator across all possible ***samples*** (the estimator's expected value) equals the parameter being estimated, the estimator is said to be unbiased; otherwise, the estimator is ***biased.***

BIB (Balanced Incomplete Block) spiraling. A complex variant of ***multiple matrix sampling*** in which items are administered so that each pair of items is dispensed to a nationally representative ***sample*** of ***respondents.***

block. A group of assessment items created by dividing the item pool for an age or grade into subsets. Blocks are used in the implementation of the ***BIB spiral sample*** design.

booklet. The assessment instrument created by combining ***blocks*** of assessment items.

calibrate. To estimate the parameters of a set of items using responses of a ***sample*** of examinees.

clustering. The process of forming sampling units as groups of other units.

codebook. A formatted printout of NAEP data for a particular ***sample*** of ***respondents.***

coefficient of variation. The ratio of the standard deviation of an estimate to the value of the estimate.

common block. A group of background items included at the beginning of every assessment ***booklet.***

conditional probability. Probability of an event happening, given the occurrence of another event.

conditioning variables. Demographic and other background variables characterizing a ***respondent.*** These variables are used to construct plausible values.

constructed-response item. A non-multiple-choice item that requires some type of written or oral response.

degrees of freedom [of a variance estimator]. The number of independent pieces of information used to generate a ***variance*** estimate.

derived variables. Subgroup data that were obtained through interpretation, classification, or calculation procedures rather than from assessment responses.

design effects. The ratio of the ***variance*** for the ***sample*** design to the variance for a ***simple random sample*** of the same size.

distractor. An incorrect response choice included in a multiple-choice item.

excluded student questionnaire. An instrument completed for every student who was selected to participate but ultimately excluded from the assessment.

excluded students. Sampled students determined by the school to be unable to participate because they have limited English language proficiency or a disability.

expected value. The average of the ***sample*** estimates given by an estimator across all possible ***samples.*** If the estimator is ***unbiased,*** then its expected value will equal the population value being estimated.

field test. A pretest of items to obtain information regarding clarity, difficulty levels, timing, feasibility, and special administrative situations. The field test is performed before revising and selecting the items to be used in the assessment.

focused BIB spiraling. A variation of ***BIB spiraling*** in which items are administered so that each pair of items within a subject area is dispensed to a nationally representative ***sample*** of ***respondents.***

foils. The correct and incorrect response choices included in a multiple-choice item.

group effect. The difference between the mean for a specific group and the mean for the nation.

imputation. Prediction of a missing value based on some procedure, using a mathematical model in combination with available information. See ***plausible values.***

imputed race/ethnicity. The race or ethnicity of an assessed student as derived from his or her responses to particular common background items. A NAEP ***reporting subgroup.***

item response theory (IRT). Test analysis procedures that assume a mathematical model for the probability that a given examinee will respond correctly to a given exercise.

jackknife. A procedure that estimates ***standard errors*** of percentages and other statistics. It is particularly suited to complex ***sample*** designs.

machine-readable catalog. Computer-processing control information, ***IRT*** parameters, ***foil*** codes, and labels in a computer-readable format.

metropolitan statistical area (MSA). An area defined by the federal government for the purposes of presenting general-purpose statistics for metropolitan areas. Typically, an MSA contains a city with a population of at least 50,000 and includes its adjacent areas.

multiple matrix sampling. Sampling plan in which different ***samples*** of ***respondents*** take different ***samples*** of items.

multistage sample design. Indicates more than one stage of sampling. The following is an example of three-stage sampling: (1) ***sample*** of counties (***primary sampling units*** or ***PSUs***), (2) ***sample*** of schools within each ***sample*** county, and (3) ***sample*** of students within each ***sample*** school.

NAEP scales. The scales common across age or grade levels and assessment years used to report NAEP results.

nonresponse. The failure to obtain responses or measurements for all ***sample*** elements.

nonsampling error. A general term applying to all sources of error, with the exception of ***sampling error.*** Includes errors from defects in the ***sampling frame,*** response or measurement errors, and mistakes in processing the data.

objective. A desirable education goal accepted by scholars in the field, educators, and concerned laypersons and established through a consensus approach.

observed race/ethnicity. Race or ethnicity of an assessed student as perceived by the exercise administrator.

oversampling. Deliberately sampling a portion of the population at a higher rate than the remainder of the population.

parental education. The level of education of the mother and father of an assessed student as derived from the student's response to two assessment items. It is a NAEP ***reporting subgroup.***

percent correct. The percentage of a target population that would answer a particular exercise correctly.

plausible values. Proficiency values drawn at random from a conditional distribution of a NAEP ***respondent,*** given his or her response to cognitive exercises and a specified subset of background variables (***conditioning variables***). The selection of a ***plausible value*** is a form of ***imputation.***

poststratification. Classification and weighting to correspond to external values of selected sampling units by a set of strata definitions after the ***sample*** has been selected.

primary sampling unit (PSU). The basic geographic sampling unit for NAEP. Can be either a single county or a set of contiguous counties.

probability sample. A ***sample*** in which every element of the population has a known, nonzero probability of being selected.

pseudoreplicate. The value of a statistic based on an altered ***sample.*** Used by the ***jackknife variance*** estimator.

random variable. A variable that takes on any value of a specified set with a particular probability.

region. One of four geographic areas used in gathering and reporting data: Northeast, Southeast, Central, and West (as defined by the Office of Business Economics, U.S. Department of Commerce). A NAEP ***reporting subgroup.***

reporting subgroup. Groups within the national population for which NAEP data are reported; for example, gender, race/ethnicity, grade, age, level of ***parental education, region,*** and ***type of location.***

respondent. A person who is eligible for NAEP, is in the ***sample,*** and responds by completing one or more items in an assessment booklet.

response options. In a multiple-choice question, alternatives that can be selected by a ***respondent.***

sample. A portion of a population, or a subset from a set of units, that is selected by some probability mechanism for the purpose of investigating the properties of the population. NAEP does not assess an entire population but rather selects a representative ***sample*** from the group to answer assessment items.

sampling error. The error in survey estimates that occurs because only a ***sample*** of the population is observed. Measured by sampling ***standard error.***

sampling frame. The list of sampling units from which the ***sample*** is selected.

sampling weight. A multiplicative factor equal to the reciprocal of the probability of a ***respondent*** being selected for assessment with adjustment for ***nonresponse*** and, perhaps, ***poststratification***. The sum of the weights provides an estimate of the number of persons in the population represented by a ***respondent*** in the sample.

school questionnaire. A questionnaire completed for each school by the principal or other official. It is used to gather information concerning school administration, staffing patterns, curriculum, and student services.

secondary-use data files. Computer files containing respondent-level cognitive, demographic, and background data. They are available for use by researchers wishing to perform analyses of NAEP data.

selection probability. The chance that a particular sampling unit has of being selected in the ***sample.***

session. A group of students reporting for the administration of an assessment. Most schools conduct only one session, but some large schools conduct as many as 10 or more.

simple random sample. The process for selecting n sampling units from a population of *N* sampling units so that each sampling unit has an equal chance of being in the ***sample*** and every combination of *n* sampling units has the same chance of being in the ***sample*** chosen.

standard error. A measure of sampling variability and measurement error for a statistic. Because of NAEP's ***complex sample design,*** sampling standard errors are estimated by ***jackknifing*** the ***samples*** from first-stage ***sample*** estimates. Standard errors may also include a component due to the error of measurement of individual scores estimated using ***plausible values.***

stratification. The division of a population into parts, or strata.

stratified sample. A ***sample*** selected from a population that has been stratified, with a ***sample*** selected independently in each stratum. The strata are defined for the purpose of reducing ***sampling error.***

student ID number. A unique identification number assigned to each ***respondent*** to preserve his or her anonymity. NAEP does not record the names of any ***respondents.***

subject area. One of the areas assessed by NAEP; for example, art, civics, computer competence, geography, literature, mathematics, music, reading, science, U.S. history, or writing.

systematic sample (systematic random sample). A ***sample*** selected by a systematic method; for example, units selected from a list at equally spaced intervals.

teacher questionnaire. A questionnaire completed by selected teachers of ***sample*** students. It is used to gather information concerning years of teaching experience, frequency of assignments, use of teaching materials, and availability and use of computers.

Trial State Assessment Program. A NAEP program authorized by Congress in 1988 and established to provide for a program of voluntary state-by-state assessments on a trial basis.

trimming. A process by which extreme weights are reduced (trimmed) to diminish the effect of extreme values on estimates and estimated ***variances.***

type of location (TOL). One of the NAEP ***reporting subgroups,*** dividing the communities in the nation into groups based on the proportion of the students living in each of three sizes and types of communities.

variance. The average of the squared deviations of a ***random variable*** from the expected value of the variable. The variance of an estimate is the squared ***standard error*** of the estimate.

A

Academic researchers, and confidentiality issues, 13–14
Achievement levels, 27–29
Advanced Studies Seminar on Use of NAEP Database for Research and Policy Discussion, 48
American College Testing (ACT), 1
American Institutes for Research (AIR), 1
Arts assessment (1997), 25–26, 40
  framework for, 8–9
Assessment clearance package, 21
Assessment development process, 20–22
Assessment innovations, 25–26
  impact of, 26
Assessment schedules, 54–56

B

Background questionnaires, 5, 13, 27, 34, 42–44
  development of, 20–22
  parent access to, 15–17
  results provided by, 27
Balanced incomplete block (BIB) spiraling, 32, 56–59
Bilingual booklets, 23
Blocks, 21–22, 42, 57–58

C

Calibration scoring, 33, 61
CCSSO. *See* Council of Chief State School Officers
Center for Research on Evaluation, Standards, and Student Testing (CRESST), 17
Civics assessment (1998), 30
Commissioner of Education Statistics, 1, 17
Comparability of state and national results, 38–39
Computer scanning of response booklets, 30, 32, 61
Confidentiality of assessment data, 13–14, 61
Congressional mandate for assessment evaluation, 17–19
Constructed-response questions, 4
  and assessment scales, 28
  number of, 7
  in 1998 reading assessment, 26
  scoring of, 30–31
  validity and evaluation of, 18
Contextual background variables, 27–28, 42–44. *See also* Background questionnaires
Council of Chief State School Officers (CCSSO), 1
CRESST. *See* Center for Research on Evaluation, Standards, and Student Testing
Cross-State Data Compendia, 41

D

Data
  collection of, 54, 60–61
  confidentiality of, 13–14, 61
  international linking of, 49–50
  secondary analysis of, 47–48
Data collection procedures, 54, 60–61
Data products, 47
Demonstration booklet, 15–16
Department of Education
  Freedom of Information Act officer, 16
  Information Management Compliance Division (IMCD), 21–22
Destruction Notice, 14
Differential item functioning (DIF), 22, 35
Disabilities, students with
  accommodation for, 23–24
  background questionnaires for, 42, 44
Documentation of scoring process, 33

E

Educational research, assistance for, 47–48
Educational Testing Service (ETS), 1, 8, 20
  scoring and processing by, 30–33
  software for secondary data analysis, 48
  technical assistance provided by, 47
  test materials produced by, 61
Educators, resources for, 45–46
Electronic access to reports, 40, 47
Estimation, 34–36
ETS. *See* Educational Testing Service
Evaluation of assessments, 17–19
Excluded students, 23–24
Expert panels, 17–19

F

Field-test clearance package, 21–22
Field tests, 22
Focused Reports, 41
Frameworks
  development of, 3, 6–8
  educational uses of, 45–46
  for 1999 long-term trend assessment, 8
  for 2000 mathematics assessment, 8–9
  for 2000 reading assessment, 11–12
  for 2000 science assessment, 9–10

Freedom of Information Act officer (Department of Education), 16

G

Grade levels included in 1999–2000 assessments, 6–7, 54
Group-level statistics, 34

I

IAEP. *See* International Assessment of Educational Progress
Image-based scoring system, 30, 32, 61
Improving America's Schools Act of 1994, Title IV, 13
Inclusion, increased, 23, 44
Individualized Education Plans (IEPs), 23, 44
Information Management Compliance Division (IMCD), Department of Education, 21–22
Instructional experiences, assessment of, 27
Instructional reports, 40
Instrument Development Committees, 20–22
Instruments
  development of, 20–22
  flexibility of, 4
International assessment data, link to, 49–50
International Assessment of Educational Progress (IAEP), 49
Interpreting, 8
Interrater reliability, 32
Item analysis, 34–35
Item response theory (IRT), 28, 34–36

L

Limited English Proficiency (LEP)
  accommodation for, 34
  background questionnaires for, 42, 44, 61
Linking, 6, 17, 36, 39, 49–50
Long-term trend assessments, 4, 7–8
  data collection procedures, 60
  framework for, 8
  frequency and timing of, 54–56
  student selection for, 51–52

M

Main assessments, 4. *See also* National assessments
Maryland Learning Outcomes, 26
Mathematics assessments, 25, 35, 38, 40, 54
  framework for, 8–9
Matrix sampling, 57–59
Multiple-choice questions
  and assessment scales, 28, 30
  scoring of, 28, 30
  vs. constructed-response questions, 7–8

N

NAEP. *See* National Assessment of Educational Progress
NAEPEX, 48
NAEP NETWORK, 8, 21–22, 48
National Academy of Education (NAE), Panel on the Evaluation of NAEP Trial State Assessment Project, 18
National Academy of Sciences (NAS), 19
National Assessment Governing Board (NAGB), 1
  achievement levels, 28–29
  framework development by, 4, 6–8
  instrument development by, 20–22
  scoring guide review by, 31
  subject selection by, 6
National Assessment of Educational Progress (NAEP) 1996, 40
  1997, 6–8, 25–26
  1998, 6–7, 9–12, 26, 38–41
  comparability of results, 38–39
  frequency and timing of, 3, 54
  goals of, 3
  legislation authorizing, 13, 18
  long-term trend. *See* Long-term trend assessments
  national. *See* National assessments
  overview of, 1, 3–6
  state. *See* State assessments
National assessments, 3–4
  1998, sample characteristics of, 54
  1999, subject areas and grades included in, 6–8
  2000, subject areas and grades included in, 6–12
  analysis of, 34
  background questionnaires for, 42–44
  comparability to state assessments, 38–39
  development of, 20–22
  frequency and timing of, 54
  number of schools and students included in, 54–56
  results of, 27
  scaling of, 34–36
  student selection for, 51–52
  validity and evaluation of, 17–19

National Center for Education Statistics (NCES), 1–2
data collection operations, 60–61
expert panels established by, 17
framework development by, 6–8
increased inclusion policies, 23
instrument development by, 20–22
reports published by, 13
scoring guide review by, 30–31
technical assistance provided by, 47–48
test booklets released for review by, 15
Web site, 40, 46, 47
National Computer Systems (NCS), 1
scoring and processing by, 30–32
test materials produced by, 61
National Education Statistics Act of 1994, 13
National Education Summit, 49
Nonbias policy, 22
Noncognitive blocks, 42
North Carolina End-of-Grade Tests in Mathematics, 26
NVS Panel. *See* Validation Studies Panel

O

Office of Management and Budget (OMB), 21–22
Ohio state assessments, 26
Oversampling, 53

P

Parent access to test booklets, 15–16
Parental consent, 13
Participation in NAEP, 3–5, 13, 23, 54
Plausible values methodology, 36–37
Primary sampling units (PSUs), 52
Public Law 100-297, 18
Public Law 103-382, 18
Public release of questions, 15–16

Q

Quality control measures
for data collection operations, 60–61
to monitor against bias, 22
for scaling and analysis of data, 34–36
for scoring process, 28–29
Questionnaires, background. *See* Background questionnaires
Questions
constructed-response. *See* Constructed-response questions
multiple-choice. *See* Multiple-choice questions
parent access to, 15–16
released, 15–16, 45–46
sample, 16
security of, 15–16

R

RAND, 17
Reading
for information, 9–11, 28
for literary experience, 9–11, 28
to perform tasks, 9–11, 28
Reading assessment
1994, scales developed for, 28–29
1998, 9–11, 26, 38–39, 54
framework for, 9–10
sample questions from, 9–11
Reading composite scale, 28
Reading literacy, aspects of, 9–11
Reading situations, types of, 9–10, 28
Reading stances, 9–10
Report Cards (NAEP), 40
Reports, 40–41, 46
educational uses of, 45–46
focused, 41
instructional, 40
state, 40
technical, 41
trend, 41
update, 40
Results, 27–29
analysis of, 34–37
processing and scoring of, 30–33

S

Samples, student, 51–53
Sample selection, 54–55
Sampling, 54, 57–59
SAS. *See* State aggregate sample
Scales, linking of, 36
Scale-score data files, generation of, 35–36
Scaling, 34–36, 37
of national and state assessments, 38–39
School environment, assessment of, 27
School questionnaires, 42–44. *See also* Background questionnaires
Schools
number of, 54–56
selection of, 52
Science assessments, 25, 38, 54
framework for, 9–10
Scorers
consistency/monitoring of, 30, 32–33
recruiting and training of, 30–32
Scoring
of constructed-response questions, 45
documentation of, 33, 61
reliability of, 30–33, 61
of state assessments, 31
for trend questions, 33
Scoring guides, 30–31
Scoring system, image-based, 30, 32, 61
SCS. *See* State comparison sample

SD. *See* Students, with disabilities
Security, importance of, 15–16
Seminars, 48
Sensitivity review, mandatory, 22
Software products, 41, 47
Special studies/analyses, 48
Spiraling, focused BIB, 57–59
State aggregate sample (SAS), 38–39
State assessments, 3–5
  1998, sample characteristics of, 54
  2000, subject areas and grades included in, 6–8, 54–55
  analysis of, 34
  background questionnaires for, 42
  comparability among, 38–39
  comparability with national assessments, 38–39
  development of, 20–22
  frequency and timing of, 54
  mandated, 13
  number of schools and students included in, 54
  questions included in, 8
  results provided by, 27
  scaling of, 38–39
  scoring of, 31
  student selection for, 51–53
  validity and evaluation of, 18–19
State comparison sample (SCS), 38–39
State-level forum, for educational issues and policy discussions, 48
State Reports, 40–41
Stratification, 52–53
Student achievement
  estimation of, 34–36
  trends in, measurement of, 32–33, 35–36
Student motivation questions, 43
Student questionnaires, 42–43. *See also* Background questionnaires
Students
  with disabilities
    accommodation for, 23–24
    background questionnaires for, 42, 44
  excluded, 23–24
  with limited English proficiency
    accommodation for, 23–24
    background questionnaires for, 42, 44
  number of, 54–56
  populations of, assessment of, 27
  sampling, 51–53
  selection of, 51–53
Subject area questions, on student questionnaires, 43
Subject-matter achievement, assessment of, 27–29
Subjects
  included in 2000 assessments, 6–7, 9–12
  selection of, 6
Subject scales, 28
Summary Data Tables Tool, 41, 48

T

Teacher questionnaires, 43. *See also* Background questionnaires
Technical assistance, 47–48
Technical Reports, 41
Technical Review Panel (TRP), 17
Test booklets
  bilingual, 23
  computer scanning of, 30–31, 61
  numbering of, 60
  parent access to, 15–16
  planned variation of, 57–59
  warehousing of, 33
Test materials, production and management of, 60–61
Testing time, extended, 23
Third International Mathematics and Science Study (TIMSS), 49–50
Training
  of data collection administrators, 60
  of scorers, 30–32
Trend Reports, 41
Trial State Assessment (TSA), 4–5, 18
TRP. *See* Technical Review Panel

U

University of Colorado at Boulder, 17
Update reports, 40
Using NAEP data, 42–50

V

Validation Studies Panel, 18
Validity, of assessments, 17–19
Voluntary nature, of assessments,

W

Weighting, 34, 53
Westat, 60–61
World Wide Web site (NCES), 40, 45–47
Writing assessment (1998), 30

Subject Areas Assessed by NAEP

Year	NATIONAL NAEP		STATE NAEP	
	Main Assessment	Long-Term Trend		
2000	Mathematics, Science, Reading (Grade 4)		Mathematics Science	Grades 4, 8 Grades 4, 8
1999		Mathematics, Science, Reading, Writing		
1998	Civics, Reading, Writing		Reading Writing	Grades 4, 8 Grade 8
1997	Arts (Grade 8)			
1996	Mathematics, Science	Mathematics, Science, Reading, Writing	Mathematics Science	Grades 4, 8 Grade 8*
1994	Geography, U.S. History, Reading	Mathematics, Science, Reading, Writing	Reading[3]	Grade 4
1992	Mathematics, Reading, Writing	Mathematics, Science, Reading, Writing	Mathematics[4] Reading[4]	Grades 4, 8 Grade 4
1990	Mathematics, Science, Reading	Mathematics, Science, Reading, Writing	Mathematics[5]	Grade 8
1988	Civics, Document Literacy,[2] Geography,[2] U.S. History, Reading, Writing	Civics[3] Mathematics, Science, Reading, Writing	*Since 1994, states have assessed nonpublic schools as well as public schools.*	
1986	Computer Competence, U.S. History,[2] Literature,[2] Mathematics, Science, Reading	Mathematics, Science, Reading[4]	*Department of Defense schools were assessed at Grades 4 and 8.	
1984	Reading, Writing	Reading, Writing		
1981–82	Mathematics, Science, Citizenship/ Social Studies[3]	Mathematics,[3] Science[3]		
1979–80	Reading/Literature, Art	Reading[3]		
1978–79	Art, Music, Writing			
1977–78	Consumer Skills,[2] Mathematics	Mathematics[3]		
1976–77	Basic Life Skills,[2] Science	Science[3]		
1975–76	Citizenship/Social Studies, Mathematics[3]	Citizenship/Social Studies[3]		
1974–75	Art, Index of Basic Skills, Reading	Reading[3]		
1973–74	Career and Occupational Development, Writing			
1972–73	Mathematics, Science	Mathematics,[3] Science[3]		
1971–72	Music, Social Studies			
1970–71	Literature, Reading	Reading[3]		
1969–70	Citizenship, Science, Writing	Science[3]		

Explanation of format for Year column: Before 1984, the main NAEP assessments were administered in the fall of one year through the spring of the next. Beginning with 1984, the main NAEP was administered after the new year in the winter, although the assessments to measure long-term trend continued with their traditional administration in fall, winter, and spring. Because the main assessment is the largest component of NAEP, beginning with 1984 we have listed its administration year rather than the two years over which trend continued to be administered. Note also that the state component is administered at essentially the same time as the main NAEP.

[2] This was a small, special-interest assessment administered to limited national samples at specific grades or ages and was not part of a main assessment. Note that this chart includes only assessments administered to in-school samples; not shown are several special NAEP assessments of adults.

[3] This assessment appears in reports as part of long-term trend. Note that the Civics assessment in 1988 is the third point in trend with Citizenship/Social Studies in 1981–82 and in 1975–76. There are no points on the trend line for Writing before 1984.

[4] The 1986 long-term trend Reading assessment is not included on the trend line in reports because the results for this assessment were unusual. Further information on this reading anomaly is available in Beaton and Zwick (1990).

[5] State assessments in 1990–94 were referred to as Trial State Assessments (TSAs).

BEST COPY AVAILABLE

United States
Department of Education
Washington, DC 20208–5653

Official Business
Penalty for Private Use, $300

Presorted Standard
POSTAGE & FEES PAID
U.S. Department of Education
Permit No. G–17

Standard Mail (A)